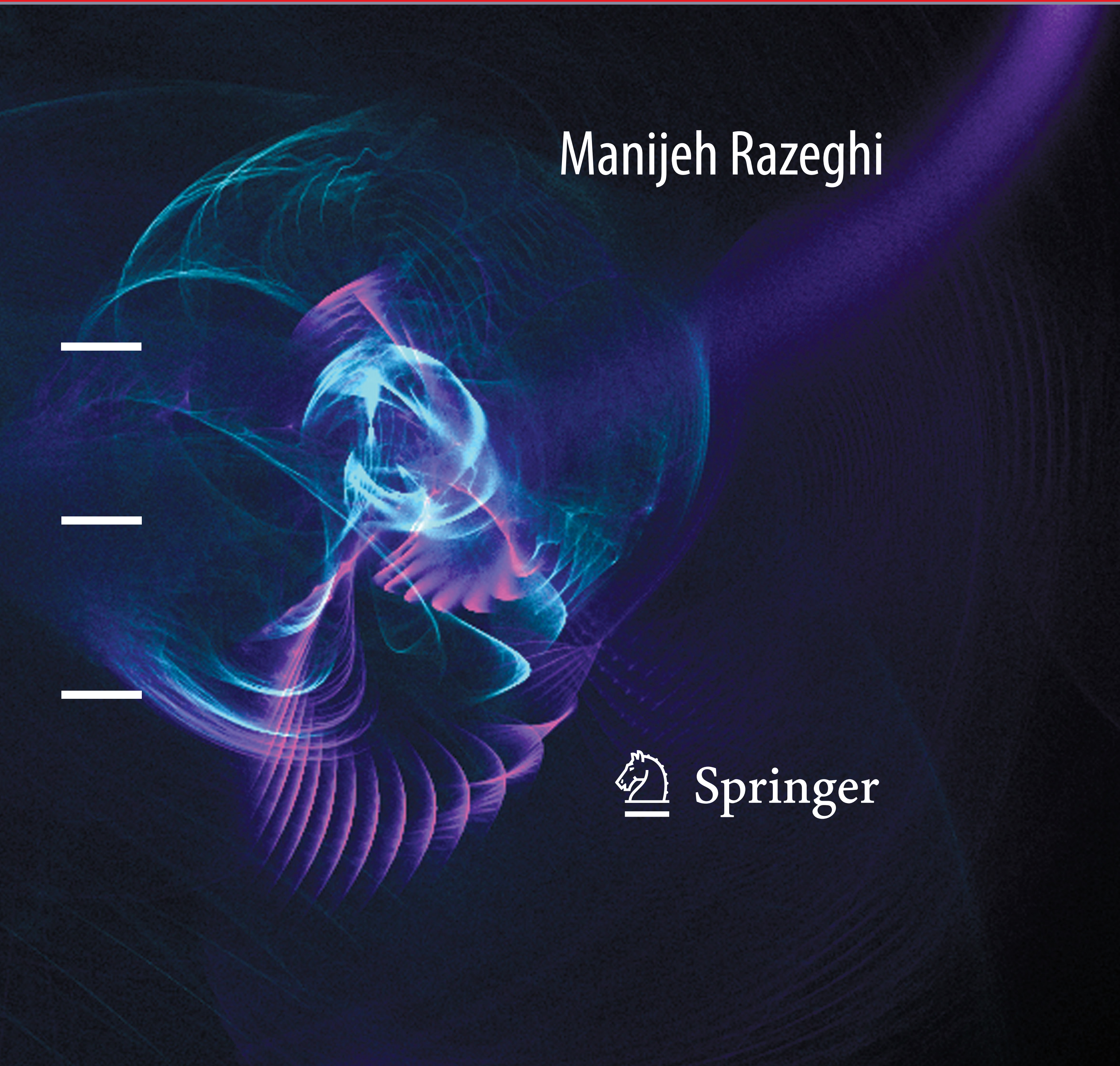
Technology of Quantum Devices

Manijeh Razeghi



Springer

Technology of Quantum Devices

Manijeh Razeghi

Technology of Quantum Devices

Springer

Manijeh Razeghi
Walter P. Murphy Professor of Electrical Engineering
and Computer Science
Northwestern University
2220 Campus Dr. RM 4051
Evanston, IL 60208–3129
USA

ISBN 978-1-4419-1055-4 e-ISBN 978-1-4419-1056-1
DOI 10.1007/978-1-4419-1056-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009935032

© Springer Science+Business Media, LLC 2010
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Students commonly think of a textbook as merely a tool to get prepared for exams. This is not the right way of looking at it! A textbook is the fruit of long-term studies and experience acquired by the author and reflects her or his personality. It embodies priorities; knowledge and I dare say even dreams and life attitudes. Compare the difference in style and content in the now classic physics textbooks by Landau and Feynman. Both Landau and Feynman were scientists whose minds were ready to listen to the music of the heavens. But how very differently! Landau wrote with the authority of a Zeus and his book sounds like the ultimate message from Heaven, while Feynman's style is more modest, and his curiosity and quest for truth could hardly be matched by anyone. His famous textbook is like an invitation to travel through the Disneyland of Nature, where he acts as a guide, but a guide who is also learning during this journey. And there is a third example: the Chicago lecture notes on quantum mechanics by another Nobel laureate – Enrico Fermi. At first sight – it appears to be more student friendly, simple, very much to the point, but what a simplistic, and, indeed, incorrect interpretation that would be! Fermi made a selection of topics and then reduced the content to the absolute essence of what has to be understood to get prepared for a journey into the quantum wonderland. He did it in such a way that an average student had the impression he or she understood everything, while a more demanding student would get a sense of much more: a feeling that a miraculous quantum world was waiting for him behind invisible doors, full of questions and surprises. Fermi did what Albert Einstein once said about science in his peculiar English – *make things as simple as possible, but no simpler.*

I admire this textbook by Professor Razeghi as much as I respect her research achievements, which she fulfilled in her personal journey through this demanding life. She was born in Persia, but left her motherland forever to join her new country France, the country that gave her the chance to continue the science she loved so much. In doing so, she followed the footsteps of Marie Curie, who a century before left oppressed Poland as a young math-teacher by the name of Skłodowska. Welcomed in France, Skłodowska completed her studies at the Sorbonne, got married to a brilliant French physicist Pierre Curie, and then spent endless hours working with him, processing tons of radioactive ores from Czechoslovakia. Together they, eventually extracted small grains of the miraculous Polonium and Radium – two radioactive elements they discovered and named. This superb technological achievement, of which Marie definitely was the master and the *spiritus movens*, opened new avenues for science and finally led her twice to Stockholm to be awarded the Nobel medal.

Dr Razeghi hopefully was not forced to work in a cold and primitive warehouse, like the Curies had to. The wise management of the French electronic giant Thomson spotted her unique talents and gave her proper resources to realize her visions and dreams. In a short time she became the First Lady in solid-state physics and made Thomson the leader in modern III-V compound semiconductor technology. Her laboratory was a dream for most of us, well before the common excellence of today in many places. But Razeghi became a technologist by choice. She was driven by the vision of the ultimate device backed by a deep understanding of the science and full of curiosity. This is what guided her. No wonder she became a very desired collaborator for top labs and personalities in the semiconductor world. She soon reached the peak of the Himalayas and could well have stopped there. But not for Madame Razeghi. After many years of success, she left friendly Europe for the next grand tour of her life – to the host of most advanced material science – the United Sates. Interestingly, not to another industrial super-organization like Thomson, but to a University, where she could share her experiences, and shape the next generations. Her energy and visions attracted money, and the money helped to create one of most advanced university-based semiconductor labs in the world, visited and applauded by most Nobel laureates in the field.

So, dear reader, make sure that you learn from this book, but not only science and technology, which is presented with great clarity, skill and care (there is even an appendix how to work with dangerous chemicals in the MOCVD lab!). Maybe you will hear – just as I did – the whisper of the modestly hidden powerful message from Professor Razeghi: ***the only thing to prevent you from performing miracles in the tournament with Nature is yourself***. To win and to have pleasure, learn first, then practice in the lab, and work with your notebook. If you work hard enough and still enjoy it, you may have the stuff for the ultimate destiny – real Himalayas – the discourse with Nature: understand her laws and limitations, but also her immense and endless frontiers.

Thank you Manijeh for the guidance.

Jerzy M Langer
Professor in Physics, Institute of Physics
Polish Academy of Science, Warsaw, Poland
Fellow of the American Physical Society
Member of Academia Europaea

Preface

The cover of this book shows the beautiful interaction of two streams of cosmic dust – this serves as a philosophical allegory to the contents of this book. We start with atoms fixed in a crystalline lattice. When these atoms are of the right type, and organized correctly, they profoundly influence the behavior of electrons, similar to the cosmic dust on the cover. Arranging many atoms together creates an artificial structure within the crystal, whose electrical and optical properties are entirely within our control. By understanding the art and science of atomic engineering we can create a wide array of sophisticated semiconductor devices.

This book is dedicated to the student who is specializing in solid state engineering especially in the areas of nanotechnology, photonics, and hybrid devices. He is expected to have a basic knowledge, at undergraduate level, of the fundamentals of semiconductor physics. The present book was developed with a view to nanotechnology, which we believe is the subject of today, tomorrow being perhaps dedicated to the interface between solid state and soft solids and biology. The reader is expected to have an elementary knowledge of quantum mechanics. For example he should understand what is meant by quantum confinement and realize its novelty and importance. He is expected to have come across such concepts as "the semiconductor superlattice," "the quantum dot," "the heterojunction," and have learned why it is interesting to study these systems. In this book he is going to learn how to make devices which use the new quantum physics which results from the reduced dimensionality. (You would do well to refer to *Fundamentals of Solid State Engineering* as the ideal place to freshen up on these topics.)

To begin with, Chapter 1 of this book discusses modern single crystal semiconductor growth technology with a focus on recent development and technological improvements critical to modern semiconductor devices. In Chapter 2, we are going to learn the first steps on how to actually fabricate a bulk semiconductor device, how to prepare the material the substrate and achieve the doping. Then in Chapter 3, we consider the fabrication of an actual device structure. This involves patterning the semiconductor, and then wire bonding it to arrive at the desired circuit configuration. Patterning involves photolithography and electron beam lithographies. This is a specialized topic of great importance also for the new emerging fields of organic and hybrid electronics. So more recently, scientists and engineers have also invented the so called nano-imprint lithography in which man-sized stamps are used to impress an image onto a surface. One can now also use atomic force microscope tip to move atoms around on surfaces. “Nano-membranes” can now be fabricated by etching away the substrate and producing ultra thin free standing semiconductor films which can also be “glued” onto another surface. These new developments have given device fabrication another very powerful degree of spatial resolution and flexibility. In addition, one has to imagine the AFM (atomic force microscope) tip moving around on surfaces, and placing magnetic atoms, magnetic clusters and fluorescent molecules and nanoparticles exactly into the location where they are needed on the surface. This technology will allow us to eventually make nanomachines, tools, and even surgical instruments. It is already routine now to implant nanosized metallic particles or fluorescent molecules of engineered sizes and shapes into cancer tumors, and then to irradiate them. With metallic particle at their resonance “plasmon” absorption frequencies for example, the particles get hot and destroy the tumor with minimal damage to the rest of the tissue. The key discovery here, was that one can tune the plasmon resonance (collective oscillation frequency of the charges on the surface of the metal particle) by changing the shape and size of particles. This requires “nanoengineering” and “nanochemistry.”

Making physically contactable electrical circuits on a micron scale constitutes what is the well-established chip technology. In Chapter 4 we review the operation of the *p-n* junction which constitutes one of the fundamental building blocks of many modern electronic and opto-electronic devices. In Chapter 5 we introduce the student to the technology of the transistor. The concept of switching and amplification is explained. The various types of transistor architectures are introduced. The focus is here not so much on absolute miniaturization, as to understanding present day transistor technology. The absolute miniaturization down to single electron devices is still very much a research field. We feel that this fascinating topic should be the subject of a specialized textbook because the present book is a book for engineers. However small, the devices described in this volume are ones which have current engineering applications. In Chapter 6, we consider the principles, design and fabrication of the semiconductor laser. Later in Chapter 7 the reader will learn how one makes and operates a Quantum Cascade Laser (QCL), a work of art in the application of quantum mechanics. But first, he has to learn the principles of light amplification and light confinement, i.e. waveguiding, and how one can make lasers using semiconductors. Semiconductors lasers are ideal for the mid-infrared wavelength regime. Mid-infrared wavelengths (3–12 μm) have a remarkable amount of versatility for many new types of applications. Perhaps the most important aspect of this wavelength range is that all molecules are optically active in this regime, and quantitative infrared spectroscopy has been an industrial tool for many years. Most of the time these tools were however thermal sources, and they were limited in sensitivity and range. Mid-infrared lasers, use an excitation/illumination source, and have demonstrated very sensitive real-time and remote sensing capabilities. Besides simple spectroscopy, the direct absorption of light in the right range by specific molecules also lends itself to some potentially very useful medical technologies. Breath analysis, for example, has already been used to monitor health by checking for abnormal cell metabolism byproducts. In the future, it may also be possible to target or cauterize specific types of cells by their chemical or protein content for selective surgery. In addition, mid-infrared lasers, in some ranges, have very good atmospheric transmission, which potentially allows for improved, secure, free-space communication, which is less sensitive to weather conditions than existing near-infrared systems. The uniqueness of the QCL described in detail in Chapter 7, is that the laser transition takes place between two quantum "intersubband states", whose energy difference can be engineered to produce lasers with different wavelengths using the same material system.

In Chapter 8 we turn our attention to measuring light intensities, not creating light. Each wavelength regime has its own characteristics uses and its own applications: seeing and recording visible daylight (500–700 nm) to seeing hotter objects in the dark (2–20 μm) and or behind walls, to seeing through paper for example (THz spectroscopy). Then there is a multitude of current and potential applications for sensitive detectors in the area of specific single and multicolor detection, in the field of communication, sensing security, robotics, artificial intelligence and medical diagnostics. A sensitive photodetector is a very powerful tool, and research and development in this field is worldwide. In Chapter 8 we learn about photoconduction and how to quantify photodetector noise and define figures of merits. Then in Chapter 9, we review the most important classes of photodetectors. We explain the special role that semiconductor physics plays and how these are fabricated using single atom deposition techniques. In the current detector technologies, three examples take advantage of the low dimensional properties that are predicted by quantum mechanics. They include: the Type II InAs/GaSb superlattice photodetectors, the quantum well intersubband photodetectors QWIP and the quantum dot infrared photodetector QDIP. From the point of view of dimensionality, strictly speaking, one has to point out that the Type II superlattice is actually a three-dimensional system, the same as a bulk semiconductor, while the other two systems are respectively two and zero dimensional systems. In Chapter 10, we begin our discussion of photodetectors with Type II materials.

The concept of the Type II InAs/GaSb superlattice was first proposed by Sai-Halasz and Esaki in the 1970s. The superlattice is fabricated by alternating InAs and GaSb layers over several periods, creating a one-dimensional periodic structure, in analogy to the periodic atomic chain in naturally occurring crystals. The special feature of the Type II system is the bringing together of two materials for which the energy gaps are not aligned in energy space. The broken gap alignment as in the case InAs/GaSb leads to the situation in which electrons from the GaSb valence band can wander into the adjacent conduction band of InAs. The degree of this transfer can however be controlled by using thin InAs sandwiched layers for which the conduction band confinement can make the lowest levels again rise up above the GaSb valence band. The consequences for physics and technology are understandingly exciting. It took however a decade for this technology to reach the degree of maturity needed for the realization of the new predicted applications. Now the material systems we grow are good enough to give us the detector performance that is comparable to the state-of-the-art Mercury Cadmium Telluride (MCT) technology.

Chapter 11 is devoted to the important and beautiful area of Quantum Well (QW) and Quantum Dot (QD) physics and technology. There are several ways of fabricating small nano-size particles of semiconducting materials, but the ones we focus on in this chapter are grown using the “Stranski Krastanov” method. It was discovered by these researchers that lattice mismatch at semiconducting interfaces, could, beyond a certain point of strain, give rise to the spontaneous formation of dot like structures. The fascinating side is that these dots are fairly regularly spaced, and furthermore, they can be made to grow on top of each other. The chapter begins by introducing the basic operating principles of the intersubband detector, which are shared by the Quantum well intersubband detectors QWIPs and the Quantum Dot intersubband detectors called QDIPs. We describe how the QDIP operation deviates from the simple principles of bulk semiconductor operation when we discuss the theoretical advantages of QDIPs. Next we look at the growth technology for making the QDs that go into the QDs. The capabilities and limitations of the growth technology directly relate to whether or not the predicted theoretical advantages of QDIPs can be achieved. Finally, we finish by reviewing some of the major accomplishments in QDIP technology to date.

Whereas QDIPs and QWIPs are designed to cover the 2–15 μm range, at the other extreme, we have the UV photodetectors which operate in the <250 nm range. The high energy of the photon to be detected makes life easier, because here, we can use wide band gap materials such as the GaN, AlN and multilayers thereof materials. Large band gaps means that thermal excitation of carriers is very difficult even at room temperature, and thus the noise level is low. The growth of the GaN-AlN is however not unproblematic and the details of this exciting subject area is covered by a specialized text book devoted to these materials by the present author

{*III-Nitrides Optoelectronic Devices* by M Razeghi and M Henini, published by Elsevier 2004}.

There is another area where GaN is being usefully developed and which is causing great deal of excitement and that is the detection of single photons. This has attracted the attention of scientists now for many years. Applications include Raman spectroscopy, fluorescence spectroscopy, and importantly now also quantum computing with photons. Photons emitted by lasers can keep their coherence over long (km) distances bur the detection of the combined quantum states require the use of devices with very high level of sensitivity. Initially because of their high internal gain, photomultiplier tubes were used to demonstrate single-photon counting. However, their high volume and required voltages made these devices not so practical. Nowadays, material progress has led to the development of improved avalanche photodiodes with single-photon detection capabilities in traditional semiconductors, such as Si or InGaAs, as well as in novel wide-bandgap technologies. Integrated photon counting systems based on Si single-photon avalanche diodes (SPADs) are today commercially available for a wide spectral range from 350 nm to 900 nm; commercial InGaAs/InP avalanche photodiodes have been successfully tested as single-photon detectors at telecommunication wavelengths; and in the ultraviolet range, SiC and GaN avalanche photodiodes have demonstrated single-photon detection capabilities.

In Chapter 12, we review the basic properties of avalanche photodiodes. In the second part, we focus on the main characteristics and issues of Geiger mode operation (operating the device just above breakdown) for photon counting purposes. Towards the end of the chapter, we provide some examples of the state-of-the-art of single-photon avalanche diodes in Si, InGaAs, and GaN.

Finally in Chapter 13 we discuss the interesting developing new area of terahertz technology. In this chapter we describe recent developments in the technology of terahertz (THz) emitters. We begin, by presenting a short description of what can be done in the THz range, and later, some applications are described in detail. An overview of the different broadband sources available is presented. Further considerations are restricted to modern semiconductor THz emitters. Next, the current state of the art of the semiconductor THz emitter is presented. The Quantum Cascade Laser, which is one of the most efficient semiconductor emitters in the mid-infrared range, has now also been developed for the THz range. In this wavelength range, the device is usually fabricated using multiquantum well structures of GaAs/AlGaAs. Unfortunately these structures do not operate at room temperature under continuous wave working conditions. We conclude by discussing a new generation of THz emitters which are fabricated using the wide band-gap semiconductors GaN/AlGaN, and have the potential to allow for the realization of higher operating temperature THz QCLs.

The text is to a large extent based on original research carried out in my research group, the Center for Quantum Devices (CQD), at Northwestern University, Evanston. From the references, the reader will be able to identify the original work and he is encouraged to consult the original papers to deepen his understanding. I am grateful to my students, colleagues, and staff for their assistance during the preparation of this volume: Siamak Abdollahi-Pour, Yanbo Bai, Can Bayram, Binh-Minh Nguyen, Stanley Tsao, Dr. Shaban Darvish, Dr. Ryan McClintock, Dr. Bijan Movaghar, Dr. Jose L. Pau, Dr. Nicolas Péré-Laperne, Dr. Steven Slivken, Dr. Féréchteh H. Teherani, George Mach, and Laura Bennett. I would also like to thank Dr. Matthew Grayson for his careful reading of the manuscript and many helpful comments.

Finally I would like to express my deepest appreciation to the Northwestern University Administration for their permanent support and encouragement.

Manijeh Razeghi
Walter P. Murphy Professor of
Electrical Engineering and Computer Science

Contents

Foreword v

Preface.. vii

List of Symbols xxiii

1. Single Crystal Growth.. 1
    1.1. Introduction.. 1
    1.2. Bulk single crystal growth techniques 2
        1.2.1. Overview.. 2
        1.2.2. Czochralski techniques 3
        1.2.3. Bridgman techniques 4
    1.3. Liquid phase epitaxy 7
        1.3.1. Overview.. 7
        1.3.2. Melt epitaxy 9
        1.3.3. Liquid phase electroepitaxy (LPEE)........................... 10
    1.4. Vapor phase epitaxy (VPE).. 11
    1.5. Metalorganic chemical vapor deposition (MOCVD)........... 14
        1.5.1. Introduction.. 14
        1.5.2. MOCVD precursors 16
        1.5.3. Growth chamber designs... 18
        1.5.4. In situ characterization.. 20
    1.6. Molecular beam epitaxy (MBE)... 27
        1.6.1. Introduction.. 27
        1.6.2. Effusion cells used in MBE systems........................... 28
        1.6.3. Gas source MBE 33
        1.6.4. Metalorganic MBE.. 35
    1.7. Summary 36
    References.. 38
    Further reading... 39
    Problems 39

2. Semiconductor Device Technology 41
    2.1. Introduction... 41

2.2. Oxidation 42
  2.2.1. Oxidation process 42
  2.2.2. Modeling of oxidation 44
  2.2.3. Factors influencing oxidation rate 50
  2.2.4. Oxide thickness characterization 52
2.3. Diffusion of dopants 56
  2.3.1. Diffusion process 57
  2.3.2. Constant-source diffusion: predeposition 62
  2.3.3. Limited-source diffusion: drive-in 64
  2.3.4. Junction formation 65
2.4. Ion implantation of dopants 68
  2.4.1. Ion generation 69
  2.4.2. Parameters of ion implantation 70
  2.4.3. Ion range distribution 71
2.5. Characterization of diffused and implanted layers 74
  2.5.1. Sheet resistivity 74
  2.5.2. Junction depth 76
  2.5.3. Impurity concentration 78
2.6. Summary 79
References 80
Further reading 80
Problems 80

3. Semiconductor Device Processing 83
3.1. Introduction 84
3.2. Photolithography 84
  3.2.1. Wafer preparation 84
  3.2.2. Positive and negative photoresists 85
  3.2.3. Mask alignment and fabrication 89
  3.2.4. Exposure 91
  3.2.5. Development 92
  3.2.6. Direct patterning and lift-off techniques 93
  3.2.7. Alternative lithographic techniques 95
3.3. Electron-beam lithography 98
  3.3.1. Electron-beam lithography system 98
  3.3.2. Electron-beam lithography process 100
  3.3.3. Parameters of electron-beam lithography 102
  3.3.4. Multilayer resist systems 104
  3.3.5. Examples of structures 106
3.4. Etching 107
  3.4.1. Wet chemical etching 107
  3.4.2. Plasma etching 110
  3.4.3. Reactive ion etching 114

3.4.4. Sputter etching 114
3.4.5. Ion milling 115
3.5. Metallization 116
3.5.1. Metal interconnections 116
3.5.2. Vacuum evaporation 118
3.5.3. Sputtering deposition 121
3.6. Packaging of devices 122
3.6.1. Dicing 122
3.6.2. Wire bonding 123
3.6.3. Packaging 126
3.7. Summary 128
References 128
Further reading 128
Problems 129

4. Semiconductor *p-n* and Metal-Semiconductor Junctions 133
4.1. Introduction 133
4.2. Ideal *p-n* junction at equilibrium 134
4.2.1. Ideal p-n junction 134
4.2.2. Depletion approximation 135
4.2.3. Built-in electric field 140
4.2.4. Built-in potential 141
4.2.5. Depletion width 145
4.2.6. Energy band profile and Fermi energy 145
4.3. Non-equilibrium properties of *p-n* junctions 147
4.3.1. Forward bias: a qualitative description 148
4.3.2. Reverse bias: a qualitative description 151
4.3.3. A quantitative description 153
4.3.4. Ideal p-n junction diode equation 155
4.3.5. Minority and majority carrier currents in neutral regions 161
4.4. Metal-semiconductor junctions 163
4.4.1. Formalism 164
4.4.2. Schottky and ohmic contacts 166
4.5. Summary 169
Further reading 170
Problems 170

5. Transistors 173
5.1. Introduction 173
5.2. Overview of amplification and switching 174
5.3. Bipolar junction transistors 176
5.3.1. Principles of operation for bipolar junction transistors 177

5.3.2. Amplification process using BJTs 178
5.3.3. Electrical charge distribution and transport in BJTs... 179
5.3.4. Current gain 183
5.3.5. Typical BJT configurations.. 186
5.3.6. Deviations from the ideal BJT case 189
5.4. Heterojunction bipolar transistors 190
5.4.1. AlGaAs/GaAs HBT 191
5.4.2. GaInP/GaAs HBT 193
5.5. Field effect transistors 196
5.5.1. JFETs 196
5.5.2. JFET gate control 197
5.5.3. JFET current-voltage characteristics.......................... 198
5.5.4. MOSFETs 200
5.5.5. Deviations from the ideal MOSFET case 202
5.6. Application specific transistors 203
5.7. Summary 204
References.. 204
Problems 205

6. Semiconductor Lasers 209
6.1. Introduction.. 209
6.2. Types of lasers 210
6.3. General laser theory 211
6.3.1. Stimulated emission 212
6.3.2. Resonant cavity.. 215
6.3.3. Waveguides.. 216
6.3.4. Laser propagation and beam divergence.................... 225
6.3.5. Waveguide design considerations.............................. 228
6.4. Ruby laser 228
6.5. Semiconductor lasers 232
6.5.1. Population inversion 233
6.5.2. Threshold condition and output power 234
6.5.3. Linewidth of semiconductor laser diodes 238
6.5.4. Homojunction lasers 239
6.5.5. Heterojunction lasers 239
6.5.6. Device fabrication.. 241
6.5.7. Separate confinement and quantum well lasers 246
6.5.8. Laser packaging 249
6.5.9. Distributed feedback lasers 249
6.5.10. Material choices for common interband lasers 251
6.5.11. Interband lasers 252
6.5.12. Quantum cascade lasers 255
6.5.13. Type II lasers.. 257

  6.5.14. Vertical cavity surface emitting lasers 260
  6.5.15. Low-dimensional lasers 262
  6.5.16. Raman lasers 264
6.6. Summary 265
References 266
Further reading 268
Problems 269

7. Quantum Cascade Lasers 271
7.1. Introduction 272
7.2. Basic operation principles 273
  7.2.1. Intersubband transitions 274
  7.2.2. Cascading 275
  7.2.3. Rate equation 276
  7.2.4. Polar optical phonon resonance 283
7.3. The components of a quantum cascade laser 285
  7.3.1. Core heterostructure 285
  7.3.2. Laser waveguide 288
7.4. Making a quantum cascade laser 289
  7.4.1. Epitaxial growth and material characterization 289
  7.4.2. Processing and packaging 290
7.5. Device performance 292
  7.5.1. Power-current-voltage characteristics 292
  7.5.2. Temperature dependent characteristics 294
  7.5.3. Wall plug efficiency 296
  7.5.4. Spectra and far field 299
7.6. Wall plug efficiency optimization 300
  7.6.1. Electrical contact resistance 300
  7.6.2. Waveguide geometry 301
  7.6.3. Bonding method 304
7.7. Power scaling 306
7.8. Photonic crystal distributed feedback quantum cascade lasers ... 308
  7.8.1. Pattern design 309
  7.8.2. Coupling coefficients 311
  7.8.3. Testing results 312
7.9. Quantum cascade lasers at different wavelengths 314
  7.9.1. Short wavelength quantum cascade lasers (<4 µm) 314
  7.9.2. Mid wavelength quantum cascade lasers (4–9 µm) 315
  7.9.3. Long wavelength quantum cascade lasers (>9 µm) 315
7.10. Summary 316
References 316
Further reading 317

Problems 318

8. Photodetectors: General Concepts 321
8.1. Introduction 321
8.2. Electromagnetic radiation 323
8.3. Photodetector parameters 325
  8.3.1. Responsivity 326
  8.3.2. Noise in photodetectors 326
  8.3.3. Noise mechanisms 329
  8.3.4. Detectivity 332
  8.3.5. Detectivity limits and BLIP 333
  8.3.6. Frequency response 335
8.4. Thermal detectors 335
8.5. Summary 339
References 339
Further reading 339
Problems 339

9. Photon Detectors 343
9.1. Introduction 343
9.2. Types of photon detectors 345
  9.2.1. Photoconductive detectors 345
  9.2.2. Photovoltaic detectors 348
9.3. Examples of photon detectors 351
  9.3.1. p-i-n photodiodes 351
  9.3.2. Avalanche photodiodes 353
  9.3.3. Schottky barrier photodiodes 355
  9.3.4. Metal-semiconductor-metal photodiodes 357
  9.3.5. Type II superlattice photodetectors 357
  9.3.6. Photoelectromagnetic detectors 360
  9.3.7. Quantum well intersubband photodetectors 361
  9.3.8. Quantum dot infrared photodetectors 362
9.4. Focal Plane Arrays 363
9.5. Summary 364
References 364
Further reading 365
Problems 365

10. Type-II InAs/GaSb Superlattice Photon Detectors 367
10.1. Introduction 367
10.2. Material system and variants of Type II superlattices 368
  10.2.1. The 6.1 angstrom family 368
  10.2.2. Type II InAs/GaSb superlattice 370

    10.2.3. Variants of Sb-based superlattices 370
  10.3. Historic development of Type II superlattice photodetectors 374
  10.4. Physics of Type II InAs/GaSb Superlattices 376
    10.4.1. Qualitative description 376
    10.4.2. Quantitative calculations of electronic bandstructure 378
  10.5. Advantages of Type II superlattice 381
    10.5.1. Band gap engineering 381
    10.5.2. Auger suppression 382
    10.5.3. Large effective mass 382
    10.5.4. Normal incident, broad band absorption 383
    10.5.5. Good uniformity 384
  10.6. Material growth and characterization 385
  10.7. Device fabrication 387
    10.7.1. Single element device for testing 387
    10.7.2. Focal plane array fabrication 388
  10.8. Summary 390
  References 390
  Further reading 392
  Problems 392

11. Quantum Dot Infrared Photodetectors 395
  11.1. Introduction 396
    11.1.1. Operating principles of QWIPs and QDIPs 396
    11.1.2. Photocurrent 397
    11.1.3. Dark current 399
    11.1.4. Noise 399
  11.2. Advantages of QDIPs 400
    11.2.1. Introduction 400
    11.2.2. High gain and the phonon bottleneck 400
    11.2.3. Low dark current 401
    11.2.4. Normal incidence absorption 402
    11.2.5. Versatility 403
    11.2.6. Summary 403
  11.3. Quantum dot fabrication for QDIPs 404
    11.3.1. Introduction 404
    11.3.2. The formation of quantum dots in the Stranski-Krastanov growth mode 405
    11.3.3. Properties of Stranski-Krastanov grown dots and their effect on QDIP performance 406
    11.3.4. Quantum dot size 407
    11.3.5. Quantum dot shape 408
    11.3.6. Quantum dot density 409

  - 11.3.7. Quantum dot uniformity 410
  - 11.3.8. Conclusion and future directions for dot fabrication 412
- 11.4. Review of actual QDIP performance 412
  - 11.4.1. Introduction.. 412
  - 11.4.2. High operating temperature 412
  - 11.4.3. FPA imaging 416
  - 11.4.4. Summary 420
- 11.5. Summary 420
- References 421
- Further reading 422
- Problems 422

12. Single-Photon Avalanche Photodiodes 425
- 12.1. Introduction 425
- 12.2. Avalanche photodetectors, linear mode 427
  - 12.2.1. Device fabrication. 427
  - 12.2.2. Linear-mode operation. 429
  - 12.2.3. Excess noise. 433
- 12.3. Examples of APD structures 434
  - 12.3.1. Reach-through avalanche photodiodes. 435
  - 12.3.2. Separate absorption charge multiplication (SACM) APD 436
- 12.4. Geiger mode operation 436
  - 12.4.1. Basic theory. 436
  - 12.4.2. Passive avalanche quenching 440
  - 12.4.3. Active avalanche quenching 441
  - 12.4.4. Gated detection 442
  - 12.4.5. Device limitations 445
  - 12.4.6. After-pulsing 446
- 12.5. Examples of single-photon avalanche photodiodes 447
  - 12.5.1. Silicon single-photon avalanche diodes 447
  - 12.5.2. InGaAs/InP single-photon avalanche diodes 449
  - 12.5.3. GaN single-photon avalanche diodes 450
- 12.6. Summary 452
- References 453
- Further reading 454
- Problems 454

13. Terahertz Device Technology 457
- 13.1. Introduction 457
- 13.2. Applications 458
  - 13.2.1. THz spectroscopy 458

- 13.2.2. T-ray imaging 460
- 13.2.3. THz research tool 462
- 13.3. Broadband terahertz sources 464
- 13.4. Narrow band terahertz sources 466
  - 13.4.1. Optical converter 466
  - 13.4.2. Optically pumped gas lasers 469
  - 13.4.3. Semiconductor source based on Silicon and Germanium 469
- 13.5. Quantum cascade terahertz sources 472
  - 13.5.1. GaAs based terahertz QCLs 472
  - 13.5.2. InP based terahertz QCLs 473
- 13.6. Magnetic field effects 474
- 13.7. Difference frequency generation 480
- 13.8. GaN QCLs for high temperature operation 481
- 13.9. Summary 487
- References 488
- Further reading 493
- Problems 493

Appendices 497
- A.1. Physical constants 499
- A.2. International system of units (SI units) 501
- A.3. Physical properties of elements in the periodic table 503
- A.4. Physical properties of important semiconductors 517
- A.5. Thermionic emission 521
- A.6. Minority carrier lifetime measurement 525
- A.7. Advanced topics in Type-II photodetectors 533
- A.8. Physical properties and safety information of metalorganics 543

List of Symbols

a_0	Bohr radius
Å	Angstrom
α	Absorption coefficient
α_L	Thermal expansion coefficient
$\vec{B}$	Magnetic induction or magnetic flux density
c	Velocity of light in a vacuum
cal	Calorie
C, C_v, C_p	Heat capacity or specific heat, at constant volume, at constant pressure
d	Density
d	Distance, thickness or diameter
D, $\vec{D}$	Electric displacement
D, D_n, D_p	Diffusion coefficient or diffusivity, for electrons, for holes
$\Delta n, \Delta p$	Excess electron, hole concentration
$\vec{E}$	Electric field strength
E, E_n	Energy
E_C	Energy at the bottom of the conduction band
E_F	Fermi energy
E_{F_n}	Quasi-Fermi energy for electrons
E_{F_p}	Quasi-Fermi energy for holes
E_g	Bandgap energy
E_V	Energy at the top of the valence band
E_Y	Young's modulus
ε_0	Permittivity in vacuum
ε	Permittivity
ε_r	Dielectric constant
$F, \vec{F}$	Force
f	Frequency
f_e	Fermi-Dirac distribution for electrons
f_h	Fermi-Dirac distribution for holes
Φ_{ph}	Photon flux
Φ_B	Schottky potential barrier height
Φ_m, Φ_s	Work function of a metal, semiconductor
g	Gravitational constant
g	Density of states
G	Gibbs free-energy

G, g	Gain
Γ	Optical confinement factor
H	Enthalpy
$\vec{H}$	Magnetic field strength
h	Planck's constant
$\hbar$	Reduced Planck's constant, pronounced "*h* bar", $(=\frac{h}{2\pi})$
η	Quantum efficiency
η	Viscosity
i	$\sqrt{-1}$
i, I	Current
$J, \vec{J}$	Current density, current density vector
$J^{diff}, \overrightarrow{J^{diff}}$	Diffusion current density
$J^{drift}, \overrightarrow{J^{drift}}$	Drift current density
J_T	Thermal current
κ	Thermal conductivity coefficient
κ	Damping factor (imaginary part of the complex refractive index $\overline{N}$)
$\vec{K}$	Reciprocal lattice vector
$k, \vec{k}$	Wavenumber $(=\frac{2\pi}{\lambda}=\frac{2\pi\nu}{c})$, wavenumber vector or wavevector
k_b	Boltzmann constant
k_D	Debye wavenumber
L_n, L_p	Diffusion length for electrons, holes
λ	Wavelength
Λ	Mean free path of a particle
m, M	Mass of a particle
m_0	Electron rest mass
m^*, m_e	Electron effective mass
m_h, m_{hh}, m_{lh}	Effective mass of holes, of heavy-holes, of light holes
m_r^*	Reduced effective mass
M_V	Solid density (ratio of mass to volume)
μ	Permeability
μ_e	Electron mobility
μ_h	Hole mobility
n	Particle concentration
n	Electron concentration or electron density in the conduction band
n	Ideality factor in semiconductor junctions

$\overline{n}$	Refractive index (real part of the complex refractive index $\overline{N}$)
$\overline{N}$	Complex refractive index
N_A	Acceptor concentration
N_c	Effective conduction band density of states
N_D	Donor concentration
N_v	Effective valence band density of states
υ	Frequency
$\mathcal{N}_A$	Avogadro number
p	Hole concentration or hole density in the valence band
$p, \vec{p}$	Momentum
P	Power
Ψ	Wavefunction
q	Elementary charge
Q	Total electrical charge or total electrical charge concentration
ρ	Electrical resistivity
$\vec{r}$	Position vector
$\vec{R}$	Direct lattice vector
R	Resistance
R	Reflectivity
Ra	Rayleigh number
Re	Reynolds number
R_0	Differential resistance at V=0 bias
R_i	Current responsivity
R_v	Voltage responsivity
Ry	Rydberg constant
S	Entropy
σ	Electrical conductivity
τ	Carrier lifetime
U	Potential energy
V	Voltage
$v, \vec{v}$	Particle velocity
v_g	Group velocity
ω	Angular frequency ($= 2\pi\upsilon$)
$\vec{x}, \vec{y}, \vec{z}$	Unit vectors (Cartesian coordinates)

1. Single Crystal Growth

1.1. Introduction
1.2. Bulk single crystal growth techniques
  1.2.1. Overview
  1.2.2. Czochralski techniques
  1.2.3. Bridgman techniques
1.3. Liquid phase epitaxy (LPE)
  1.3.1. Overview
  1.3.2. Melt epitaxy
  1.3.3. Liquid phase electroepitaxy (LPEE)
1.4. Vapor phase epitaxy (VPE)
1.5. Metalorganic chemical vapor deposition (MOCVD)
  1.5.1. Introduction
  1.5.2. MOCVD precursors
  1.5.3. Growth chamber designs
  1.5.4. In situ characterization
1.6. Molecular beam epitaxy (MBE)
  1.6.1. Introduction
  1.6.2. Effusion cells used in MBE systems
  1.6.3. Gas source MBE
  1.6.4. Metalorganic MBE
1.7. Summary

1.1. Introduction

This chapter aims to provide readers with a general concept of how materials are prepared in semiconductor research and industry. High quality materials are vital to producing high quality devices. In addition, however, technologies are also in the race for cost-effective mass production with the considerations of wafer size expansion, multiple-wafer growth and compatibility with currently existing integrated circuitry.

In this chapter, we will explore recent advances of the crystal growth technologies. Starting with wafer fabrication, we will witness the technological development of the Chzochralski and Bridgman growth methods for the improvement of wafer sizes, quality uniformity and

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_1,
© Springer Science+Business Media, LLC 2010

commercial productivity. Then, epitaxial systems such as vapor phase epitaxy, liquid phase epitaxy, metalorganic chemical vapor deposition, and molecular beam epitaxy will be discussed to illustrate how reactor systems are accommodated to the specific requirements of their applications.

1.2. Bulk single crystal growth techniques

1.2.1. Overview

Despite the dominance of silicon in the wafer market, specific applications, especially in optoelectronics, require material properties that can only be offered by compound semiconductor wafers. For example, GaAs and InP substrates are used to manufacture key components for wireless and telecommunication technology, in which high mobility charge carriers are necessary for high frequency devices. For short wavelength optoelectronics, GaN is becoming a winning choice for blue and UV lasers as well as LED makers.

The most popular wafers: GaAs, InP, GaN, SiC and sapphire, now account for only 0.6% of the 8,630 million square inches annually produced in semiconductor fabrication facilities. However, that small portion is compensated by a higher cost per unit, resulting in an $800 million market size in 2007, and reaching the billion dollar threshold in 2009–2010. Besides the widely used substrates, other compound substrates such as GaSb, InAs, InSb, and CdZnTe still have a stable market position due to use in important infrared technologies based on these expensive substrates.

Regardless of substrate type, the common tasks for wafer manufacturers are to increase the wafer size, reduce the fabrication cost, and keep the same material quality. However, increase of the crystal size, increase of the charge weight, and the larger volume of the melt requires more precautions in the production process such as crucible preparation, heating methods, and pulling speed. The strong heat dissipation and the non-uniformity of the temperature profile across the crystal can also lead to the formation of defects and crystal non-uniformity. In doped substrates, controlling the doping concentration across the ingot is also a great challenge as the wafer size increases. In this section, we will describe several growth techniques which have been developed and upgraded over the last few decades to accommodate the new requirements of wafer growth.

1.2.2. Czochralski techniques

Liquid Encapsulated Czochralski (LEC) method.

The purpose of the Czochralski growth (CZ) technique is to form a cylindrical, crystalline boule. The molten precursors are kept in a temperature controlled reservoir, and a seed crystal is introduced to the top liquid interface. The seed crystal is kept at a temperature below the melting point of the crystal, which causes crystallization of the melt on the bottom face of the seed crystal. At this point, the seed crystal is pulled slowly away from the melt, and surface tension allows additional crystalline material to form underneath. Gradually, the boule grows in length, and the diameter stabilizes.

In the conventional Czochralski growth technique of III-V compounds, the III-V melt suffers from incongruent evaporation. In other words, the vapor pressure of the group V element is always higher than that of the group III element. As a result, an unprepared melt will gradually become group III rich, leading to a poor quality crystal boule. To avoid the loss of group V element, an inert liquid (normally molten Boric Oxide) is used to completely cover the melt. The Liquid Encapsulated Czochralski technique is defined by the presence of the liquid layer on the top of the molten charge. (Fig. 1.1-left)

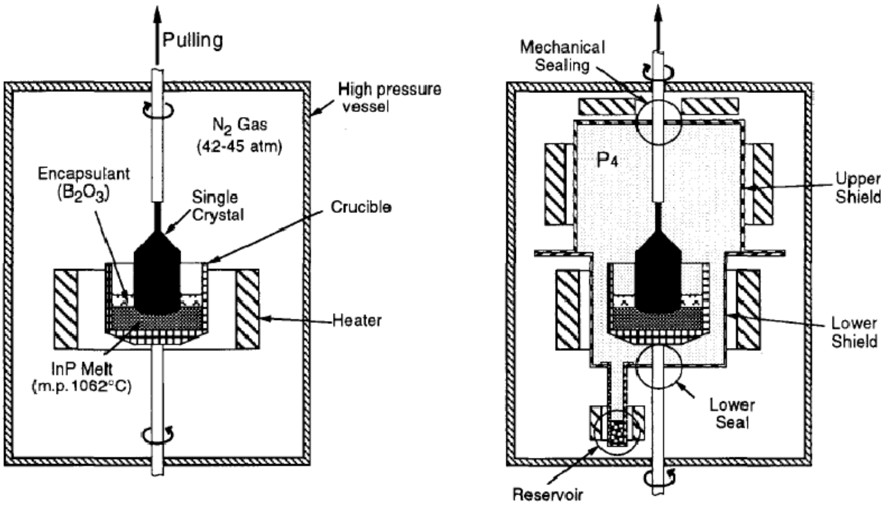


Fig. 1.1. Schematic diagram of: left) the Liquid Encapsulated Czochralski growth chamber and right) Vapor Pressure Controlled Czochralski growth chamber used in large size InP wafer growth. [Reproduced with permission from Journal of Crystal Growth Vol. 158, No. 3, Kohiro, K., Ohta, M., and Oda, O., "Growth of long-length 3 inch diameter Fe-doped InP single crystals," p. 197, Fig. 1 Copyright 1996 Elsevier Science B.V.]

Fully Encapsulated Czochralski (FEC) method:

An upgraded version of the LEC technique is the Fully Encapsulated Czochralski (FEC) method, where the whole crystal boule is fully covered by boric oxide to avoid the selective group V evaporation from the free crystal surface and to reduce the stress level at the contact boundary between the crystal and the boric oxide surfaces. This technique has been applied for 3” Si-doped GaAs substrates with good electrical properties and low dislocation density [Elliot et al. 1992]. The main problem of this method is the limitations in length and diameters of the crystals, as well as the control of the impurity content (like carbon).

Vapor pressure controlled Czochralski (VCZ) method:

The disadvantages of the conventional Czochralski and LEC techniques are the heat dissipation which induces a radial temperature gradient across the crystal boule. A powerful method to realize more homogeneous temperature fields with low temperature gradients in large wafer growth systems is the Vapor pressure controlled Czochralski (VCZ) method. The main addition is the presence of an inner chamber leading to the shielding of the growing crystal, and the heating of the atmosphere in the inner chamber (Fig. 1.1-right). In addition, to avoid group V loss at the edges of the crystal boule, a separate, heated reservoir of the group V element is used to maintain a group V overpressure. This technique has been applied for the growth of high quality 3” InP [Kohiro et al. 1996] and 4” GaAs substrates [Neubert et al. 2001]. Recently, Neubert et al. [2008] have proposed a Modified VCZ method which employs a dividing bell around the growing crystal, and they demonstrated up to 6” semi-insulating GaAs substrates.

1.2.3. Bridgman techniques

Horizontal Bridgman (HB) technique:

The Bridgman crystal growth method is similar to the Czochralski methods except for the fact that the material is completely kept inside the crucible during the entire heating and cooling processes. The seed crystal (at low temperature) is placed adjacent to the molten charge (kept at high temperature), and a lateral temperature gradient is used to control where crystallization occurs. By gradually moving the gradient away from the seed crystal as a function of time, the crystal boule grows laterally (horizontally) from the seed crystal. In order to control stoichiometry for GaAs growth, a separate source of As is heated in a separate section of the crucible, as shown in Fig. 1.2.

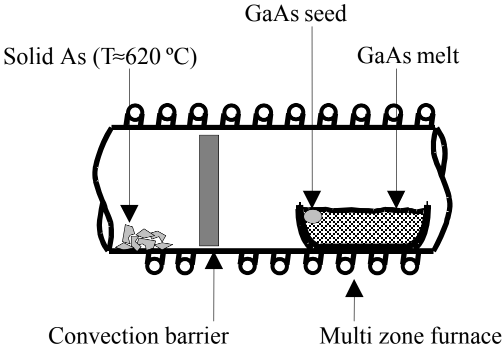


Fig. 1.2. Schematic diagram of the Bridgman growth method for a compound semiconductor such as gallium arsenide.

The first major advantage of this technique is that there is no encapsulation required, which helps avoid thermal stress at the boundary of the crystal and the encapsulation layer, and allows for a precise control of stoichiometry. The second major advantage is the relatively low temperature gradients in the solid adjacent to the solid-liquid interface, which results in a low dislocation density. However, the main disadvantages of this method are the limitation in the size of the crystals and the non-circular (D-shape) cross section that leads to difficulties and wasted material in wafer production.

Vertical Bridgman and vertical gradient freezing:

In vertical Bridgman/Vertical Gradient Freezing (VB/VGF) method, the crucible stands vertically in a furnace and contains the material charge with a seed crystal at the bottom. A temperature gradient is moved up the length of the crystal (away from the seed), promoting the single crystal growth from the seed upward. Because the crystal is formed in the shape of the crucible, the advantage of the method is the production of almost cylindrical shaped crystals up to 6” in diameter. The low thermal gradients result in low residual strains and low dislocation density. Recently, the material quality has further been enhanced by using a rotating magnetic field (RMF) in VGF-RMF equipments [O. Pätzold et al. 2004] (Fig. 1.3). The RMF affects a conducting fluid via the induced current and the Lorentz force. Under forced flow, a significant reduction of bowing of a nominally concave solid-liquid interface was demonstrated. Also, radial temperature gradients and associated thermal stresses are reduced under RMF reaction. Finally, micro segregation can also be significantly improved due to the better mixing of the melt by the RMF induced flow.

A comparison between single crystal growth techniques for GaAs was reviewed by Rudolph et al. [1999] and presented in Table 1.1. The LEC method is still the leading technology for mass production, but with reduced residual stresses and dislocation density, VCZ and/or VB/VGF will maintain a portion of the market.

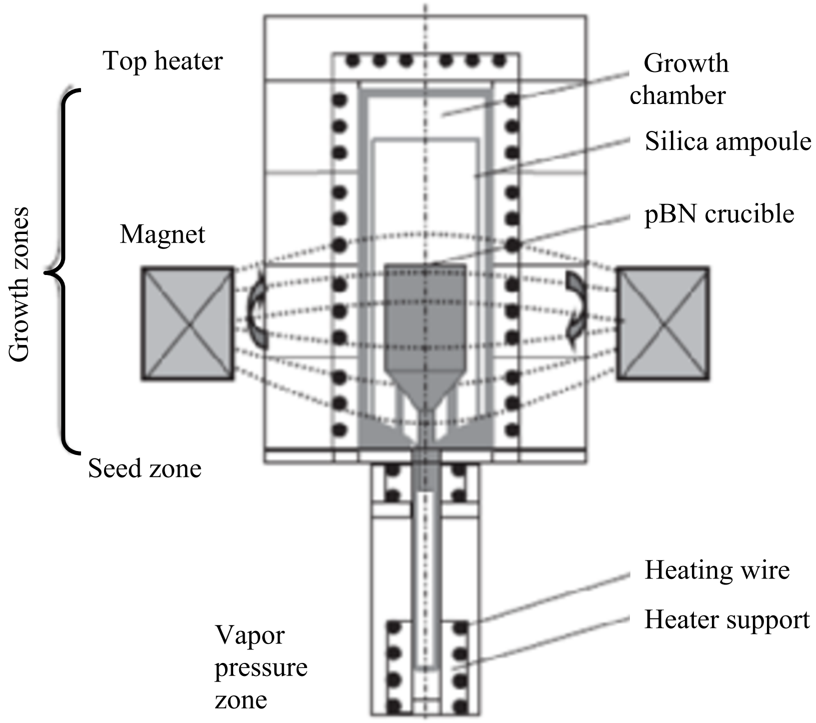

Top heater
Growth chamber
Silica ampoule
pBN crucible
Growth zones
Magnet
Seed zone
Heating wire
Heater support
Vapor pressure zone

Fig. 1.3. Schematic drawing of a VGF-RMF equipment. [Reprint with permission from Advanced Engineering Materials Vol. 6, No. 7, O. Pätzold, U. Wunderwald, M. Bellmann, P. Gumprich, E. Buhrig, and A. Cröll, "New Developments in Vertical Gradient Freeze Growth," fig. 1, pp. 554, Copyright 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim]

	LEC	VCZ	HB	VB/VGF
Scientific/technical features				
Capability for low dislocation density	Poor	Good	Very good	Very good
Uniformity of dislocation distribution	Moderate	Good	Good	Good
Cell, cell size	Small	Large	Very large	Very large
C-Control	Very good	Not solved	Poor	Not solved
Stoichiometry/EL2 control	Good	Good	Poor (?)	Very good
Crystal length	Very good	Limited	Good	Good
Crystal diameter	Very good	Very good	Limited	Very good
Background impurities	Low	Low	Low	Low
Technical realization	Good	Possible	Good	Good
Commercial features				
Investment costs	High	Very high	Low	Low
Operational costs	High	High	Medium	Low
Process maturity	Very high	Low	High	High
Productivity	High	Medium	Medium	High
Fields of application	μe-devices/ implantation	μe-devices/ implantation, LEDs	LEDs, LDs	μe-devices/ epitaxy, LEDs, LDs

Table 1.1. Comparison of industrially applied single crystal growth methods for GaAs and semiconductor compounds in general [Reprint with permission from Journal of Crystal Growth 198, Rudolph, P., and Jurisch, M., "Bulk growth of GaAs An overview," table 1, pg. 329, Copyright 1996, Elsevier Science B.V.]

1.3. Liquid phase epitaxy

1.3.1. Overview

The earliest form of epitaxy is liquid phase epitaxy (LPE), which has existed in many forms. The process relies on a thermodynamic equilibrium between a molten chemical species and the temperature-controlled solid substrate. In liquid-phase epitaxy (LPE), melts are in contact with the growing films and substrate. The precipitation of a crystalline film from a supersaturated melt onto a substrate serves as both the template for epitaxy and the physical support for the heterostructure. The early research on III-V and II-VI semiconductors was realized by liquid-phase epitaxy (LPE) (in the early 1960s) due to its simplicity and high material quality (low impurity and point defect levels). Of commercially produced devices in 1999, over 40% were GaAsP-based light emitting diodes (of these almost 50% were LPE-grown GaP devices and another 20% were LPE-grown AlGaAs/GaAs LEDs [Meyer et al. 1999])

The LPE growth technique uses a system shown in Fig. 1.4 and involves the precipitation of material from a supercooled solution onto an underlying substrate. The composition of the layers that are grown on the substrate depends mainly on the equilibrium phase diagram and to a lesser extent on the orientation of the substrate. The three parameters that can affect the growth are the melt composition, the growth temperature, and the growth time. The LPE reactor includes a horizontal furnace system and a sliding graphite boat.

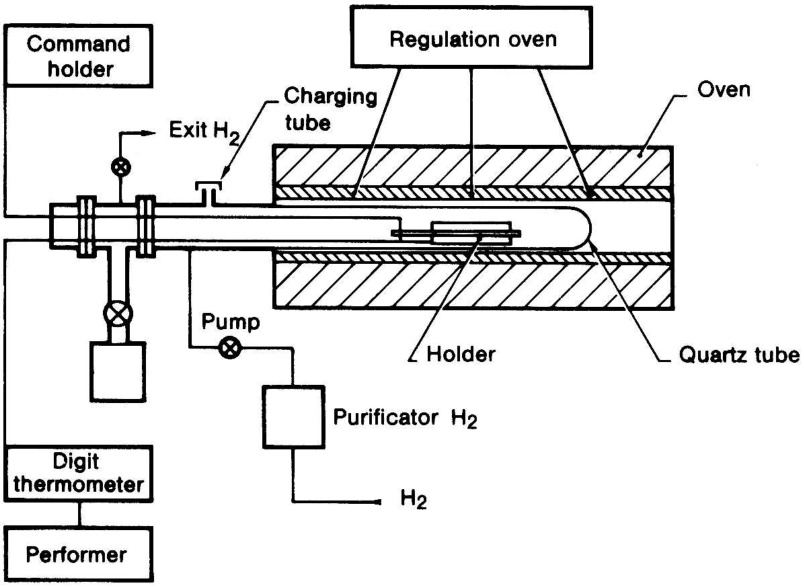


Fig. 1.4. Cross-section of a liquid phase epitaxy system. A horizontal furnace system is used with a graphite boat on which a substrate is held and which can slide inside the furnace. [Copyright © 1998 From The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications. *Fig. 1.2, p. 6. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]*

The advantages of LPE are the simplicity of the equipment used, high deposition rates, and the high purity that can be obtained. Background elemental impurities are eliminated by using high purity metals and the inherent purification process that occurs during the liquid-to-solid phase transition. The disadvantages of LPE include poor thickness uniformity, high surface roughness, melt back effects, and high growth rates which prevent the growth of multilayer structures with abrupt interfaces. Growing films as thin as a few atomic layers is therefore out of the question using liquid phase epitaxy, and is usually done using other techniques such as metalorganic chemical vapor deposition (MOCVD, discussed below).

However, in spite of these difficulties, LPE is still in use the production of simple devices such as GaAs LEDs due to the superior properties of the material produced and the cost-effective production of very thick

(10–100 μm) epitaxial layers. But, LPE is supplanted by various vapor-phase epitaxy methods (mostly MOCVD) commercially. The technique is also widely practiced to prepare epitaxial films of niobates and garnets.

1.3.2. Melt epitaxy

Recently, a new technique, called melt epitaxy (ME), has been developed by Gao et al. for the growth of low bandgap (~8–12 μm cut-off wavelength) InAsSb on GaAs and InAs substrates.[Gao et al. 2002, Gao et al. 2004] Fig. 1.5 shows the schematic diagram and growth processes on the Melt Epitaxy on a standard horizontal LPE growth system with a sliding fused boat in high-purity hydrogen ambient.

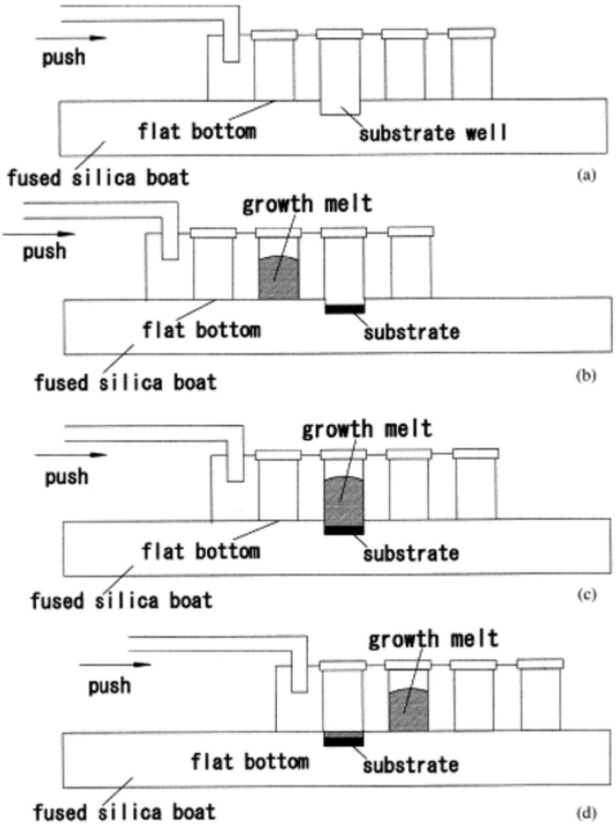

push
flat bottom
substrate well
fused silica boat
(a)
growth melt
push
flat bottom
substrate
fused silica boat
(b)
growth melt
push
flat bottom
substrate
fused silica boat
(c)
growth melt
push
flat bottom
substrate
fused silica boat
(d)

Fig. 1.5. The fused silica slide-boat schematic (a) and the slide boat arrangement before contact (b), during contact with the melt (c) and during cooling for solidification (d). [Reprinted with permission from Journal of Crystal Growth Vol. 234, No. 1, Gao, Y.Z., Kan, H., Gao, F.S., Gong, X.Y., and Yamaguchi, T., "Improved purity of long-wavelength InAsSb epilayers grown by melt epitaxy in fused silica boats," fig.1, pg. 87, Copyright 2002 Elsevier Science B.V]

Though it looks similar to standard LPE, the process is somewhat different. For LPE, the growth process is controlled by solute diffusion. Only a fraction of the melt components, dictated by the phase diagram, are deposited. As a result, the solid composition is different from the melt composition. In the case of LPE, all of the growth melt is removed from the substrate after finishing the growth. However, for ME, a portion of the pure melt is simply allowed to deposit on the substrate, and this melt fraction is crystallized under the flat part of the melt holder. As a result, the composition of the epilayer is almost same as that of the melt. With the ME technique, the layer thickness can easily be up to 100 μm.

In several reports of ME, thick InAsSb layers with cut-off wavelength between 8–12 μm were demonstrated with an electron mobility ranging from 60,000 to 100,000 $\text{cm}^2/\text{V}\cdot\text{s}$ and a background concentration of $5 \times 10^{15}\ \text{cm}^{-3}$ [Gao et al. 2002, Gao et al. 2004].

1.3.3. Liquid phase electroepitaxy (LPEE)

Liquid phase electroepitaxy, or current-controlled liquid phase epitaxy is a variant of conventional LPE where an electric current runs through the melt and the growth interface during deposition. This creates five main effects:

- The thermoelectric Peltier effect causes a vertical temperature gradient within the melt. The melt in the vicinity of the substrate is typically 1–5 °C below the melt solution temperature, resulting in supersaturation and deposition on the substrate.
- Electro-migration of the solvents in the liquid metal solutions provides a constant replenishment of the growth interface. For example, Arsenic atoms in a Ga-As melt can migrate to the growth interface and create supersaturation.
- Joule heating exists within the melt, which can counteract the Peltier effect. This can also create lateral temperature gradients which may affect uniformity and maximum wafer size.
- Increasing convection due to Joule heating in the solution results, which helps create a more uniform temperature distribution.
- Stimulation of nucleation process thanks to the supersaturation effect at the substrate/melt interface.

LPEE has been successfully employed for growth of various III-V semiconductors such as GaAs, GaP, InSb, InAs, InP and their ternaries and quaternaries. A full review of the progress of LPEE is presented by Golubev et al. [Golubev et al., 1995].

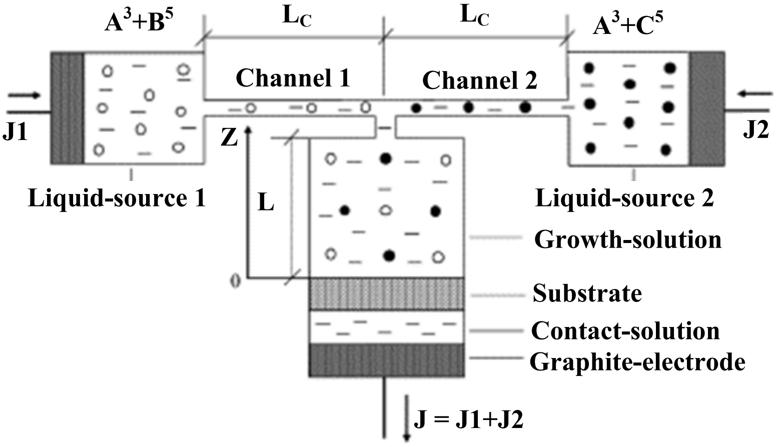


Fig. 1.6. Schematic diagram of the growth cell used for two liquid-sources LPEE of $A^3B^5_{1-x}C^5_x$*. [Reprinted with permission from Journal of Crystal Growth Vol. 249, Gevorkyan, V.A., "A new liquid-source version of liquid phase electroepitaxy," fig. 1, pg. 151 Copyright 2002, Elsevier Science B.V]*

Most recently, Gevorkyan [2003] proposed a new LPEE configurations where two liquid source are used for III-V-V ternary growths (Fig. 1.5). Epitaxial growth is performed on the A^3B^5 binary substrate using electro-transport of solute elements from two separated liquid-sources. By regulating the current from the A^3B^5 and A^3B^5 sources independently, the solid solution $A^3B^5_{1-x}C^5_x$ can be grown. The liquid-sources were separated from the growth-solution by means of sufficiently long and narrow channels 1 and 2. Modeling indicates that this technique can provide controlled compositional grading from a binary alloy composition to a target ternary alloy composition, followed by growth of a thick layer of a ternary alloy of constant composition.

1.4. Vapor phase epitaxy (VPE)

Like LPE, vapor phase epitaxy (VPE) is also a thermodynamic equilibrium growth process. However, unlike LPE, the VPE growth technique involves reactive compounds in their gaseous form. A VPE reactor typically consists of a quartz chamber composed of several zones set at different temperatures using a multi-element furnace, as illustrated in Fig. 1.7.

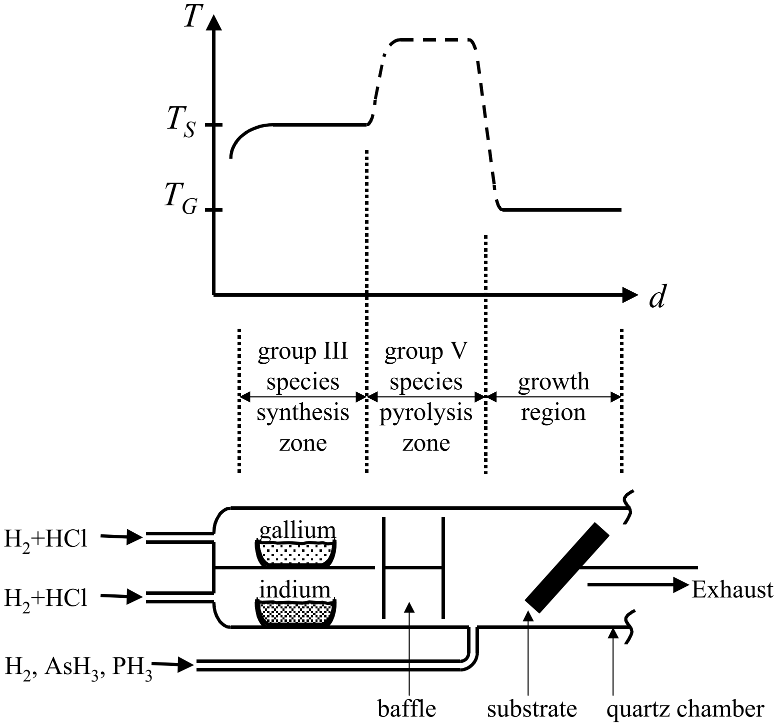

T
T_S
T_G
d
group III species synthesis zone
group V species pyrolysis zone
growth region
H_2+HCl
H_2+HCl
H_2, AsH_3, PH_3
gallium
indium
Exhaust
baffle
substrate
quartz chamber

Fig. 1.7. Cross-section schematics of a typical VPE reactor, showing the group III species synthesis, group V species pyrolysis, and the growth zones with their respective temperature profiles for the growth of a few selected semiconductors.

The group III source materials consist of pure metal elements, such as gallium (Ga) and indium (In), contained in a small vessel. In the first zone, called the group III species synthesis zone, which is maintained at a temperature T_S (~750–850 °C for GaAs or InP growth), the metal is in the liquid phase and reacts with the incoming flow of hydrogen chloride gas (HCl) in the following manner to form group III-chloride vapor compounds which can be transported to the growth region:

Eq. (1.1) $$Ga_{liq} + HCl_g \rightarrow GaCl_g + \frac{1}{2} H_{2_g}$$

Eq. (1.2) $$In_{liq} + HCl_g \rightarrow InCl_g + \frac{1}{2} H_{2_g}$$

The group V source materials are provided in the form of hydride gases, for example arsine (AsH_3) and phosphine (PH_3). In the second zone, also called the group V species pyrolysis zone, which is maintained at a temperature $T > T_S$, these hydrides are decomposed into their elemental group V constituents, yielding reactions like:

Eq. (1.3) $$AsH_3 \rightarrow \frac{u}{4}As_4 + \frac{1-u}{2}As_2 + \frac{3}{2}H_2$$

Eq. (1.4) $$PH_3 \rightarrow \frac{v}{4}P_4 + \frac{1-v}{2}P_2 + \frac{3}{2}H_2$$

where u and v represent the mole fraction of AsH_3 or PH_3 which is decomposed into As_4 or P_4, respectively.

Finally, in the growth region, which is maintained at a temperature T_G (~680–750 °C for GaAs or InP growth), the group III-chloride and the elemental group V compounds react to form the semiconductor crystal, such as GaAs or InP, on a substrate.

There are two types of chemical reactions taking place in vapor phase epitaxy, as illustrated in Fig. 1.8: heterogeneous reactions occur between a solid, liquid and/or vapor, while homogeneous reactions only occur in the gas phase.

During the growth of a semiconductor film in steady-state conditions, the overall growth process is limited by the heterogeneous reactions. During changes in the composition of the growing semiconductor, for example when switching the growth from InP to GaInAs, the process is limited by the mass transport in the gas phase.

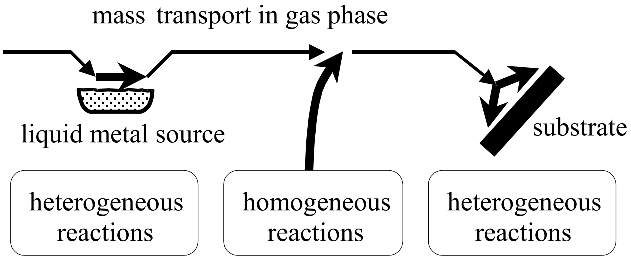


Fig. 1.8. Location of heterogeneous and homogeneous chemical reactions taking place during the vapor phase epitaxy growth process.

The advantages of VPE include a high degree of flexibility in introducing dopants into the material as well as the control of the composition gradients by accurate control of the gas flows. Growth rates are also very high, being comparable to LPE. Unlike LPE, localized epitaxy can also be achieved using VPE. A disadvantage of VPE is the difficulty to achieve multi-quantum wells or superlattices (periodic heterostructures with a large number of layers having a thickness of the order of a few tens of

Angstrom). Other disadvantages include the formation of hillocks and haze, as well as interfacial decomposition during the preheat stage.

One of the most popular uses of VPE technology today is for the fabrication of III-Nitride free-standing substrates. GaN and AlN are vital materials for next generation UV-visible optoelectronics and high power, high speed electronics. GaN, for example, is the basis of the blue laser used in the popular Blu-ray Disc^{TM} recording format.

Unfortunately, native substrate production is plagued with many problems. As a result, most of the development of this material system has been done on sapphire or silicon substrates, which have a large lattice mismatch with respect to GaN and AlN. Growth on these substrates creates crystalline defects that strongly deteriorate device performance. However, it has also been established that, as the III-Nitride layer gets thicker, there is a gradual reduction in defect density. Using VPE, which has growth rates up to 400 μm/hr, a thick layer can be produced in a short time. After removal of the original substrate, a free-standing III-Nitride substrate remains for subsequent optoelectronic or electronic device growth.

1.5. Metalorganic chemical vapor deposition (MOCVD)

1.5.1. Introduction

Metalorganic chemical vapor deposition (MOCVD) is a deposition method for thin film growth of semiconductors, metals, and ceramics. The MOCVD technology has established its ability to produce high quality epitaxial layers and sharp interfaces. This includes the growth of multilayer structures with thicknesses as thin as a few atomic layers.

A schematic diagram of an MOCVD reactor can be seen in Fig. 1.9. The growth of epitaxial layers is initiated by introducing controlled amounts of metalorganic precursors into a reaction chamber in which a heated semiconductor substrate is present. The precursors are typically volatile alkyls of group II or III elements, and either alkyls or hydrides of group V or VI elements. The decomposition of the gaseous products and reactions happens only inside the reaction chamber, leading to crystal growth.

The MOCVD system consists of four major parts: the gas handling system, the reactor chamber, the heating system and the exhaust and safety apparatus.

The gas handling system includes the alkyl and hydride sources, the valves, the pumps, and other instruments necessary to control the gas flows and mixtures. In addition, and inert carrier gas with very high purity is needed. H_2 is usually chosen for this purpose because diffusing H_2 through a palladium membrane results in a very pure gas, with particularly low levels of O_2 and H_2O. In order to minimize contamination, the gas handling system has to be clean and leak tight. In addition, the material it is made out of must be resistant to the potentially corrosive nature of the sources.

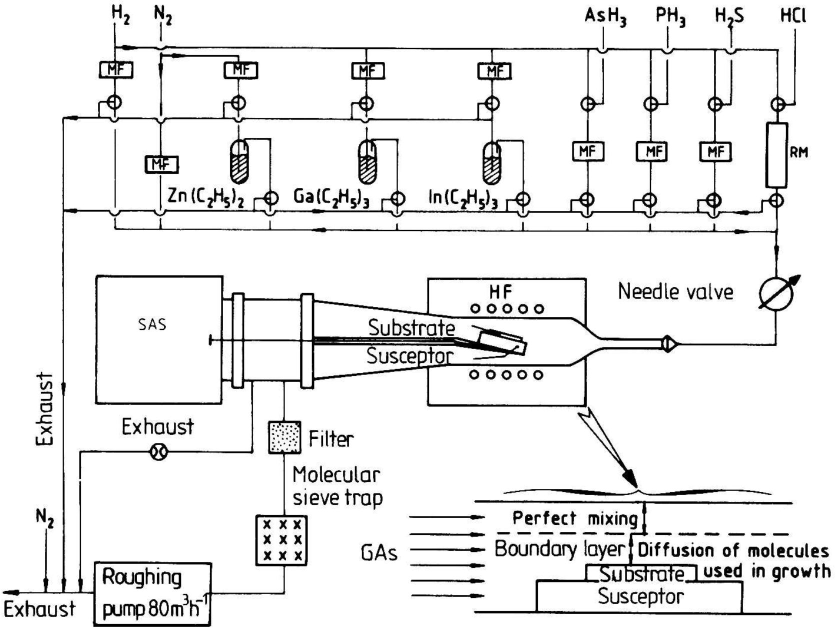

H_2
N_2
AsH_3
PH_3
H_2S
HCl
MF
RM
$Zn(C_2H_5)_2$
$Ga(C_2H_5)_3$
$In(C_2H_5)_3$
HF
Needle valve
SAS
Substrate
Susceptor
Exhaust
Exhaust
Filter
Molecular sieve trap
N_2
Exhaust
Roughing pump $80 m^3h^{-1}$
GAs
Perfect mixing
Boundary layer
Diffusion of molecules
Substrate used in growth
Susceptor

Fig. 1.9. Schematic of a MOCVD reactor

Alkyl sources are metalorganic compounds, and they are liquid or finely crushed solids, usually contained in a stainless steel cylinder called a bubbler. The purity of the sources is critical for the quality of the grown layers. As a result, much effort is constantly devoted to avoid any kind of contamination. The partial pressure of the source is regulated by precisely controlling the temperature and the total pressure inside the bubbler. Electronic mass flow controllers are used to accurately and reliably measure and control mass flow rates of the hydride and carrier gases through the gas handling system. Thus, by sending a controlled flow carrier gas through the bubbler, a controlled mass flow in the form of dilute vapors of the metalorganic compounds can be achieved.

Unlike LPE and VPE, MOCVD growth is not done in thermodynamic equilibrium. As a result, the actual growth rate is much lower than that determined from thermodynamics, because kinetics and hydrodynamic transport also play a role in determining the growth rate. Some advantages of this include enhanced doping capability, no melt back effect when changing from a low-temperature to high-temperature material, sharp interfaces, and a significantly reduced miscibility gap for most alloys.

As with most technology, MOCVD system designs are constantly improved to yield better material quality, sharper interfaces and better uniformity across the wafer. In the next sections, the sources used in MOCVD and their requirements will be discussed. Different growth chamber designs, which are able to achieve both sharp interfaces and high uniformity, will also be discussed with particular attention paid to the transition from research scale reactors to manufacturing reactors. At the end, several types in situ characterization used in MOCVD will be illustrated.

1.5.2. MOCVD precursors

Alkyls of the group II and III metals and hydrides of group V and VI elements are generally used as precursors in MOCVD. Dilute vapors of these chemicals are transported at or near room temperature to a hot zone where a pyrolysis reaction occurs. The reaction can be generalized for III-V compounds as:

Eq. (1.5) $R_3M + EH_3 \rightarrow ME + 3RH$

Where M is the group III metal(Such as Ga, In, Al), E is the group V element (such as As, P, Sb) and R is the alkyl radical (either CH_3 or C_2H_5).

The criteria for precursors in MOCVD can be summarized as:

- Saturated vapor pressure in the range of 1–10 mBar in the temperature range 0–20 °C
- Stable at room temperature and not subject to spontaneous decomposition or polymerization
- Vaporizes in H_2 without decomposing
- Reacts efficiently at the desired growth temperature
- Not subject to unwanted side reactions in the reaction chamber such as polymerization

Group III

Trimethyl and triethyl alkyls have been used exclusively in the early development of MOCVD. The trimethyl sources are most often used due to their higher vapor pressure and greater stability. TEAl, TEGa, and TEIn are only marginally stable. TEIn has been observed to decompose in storage containers. It also reacts with the group V hydrides AsH_3 and PH_3 to form non-volatile adducts in high pressure MOCVD reactors. However, in low-pressure MOCVD reactors, using TEGa and TEAl significantly reduced carbon concentrations in GaAs and AlGaAs. [Kuech et al. 1988]

Trimethylindium (TMIn) is up to now the most widely used In source. The use of TMIn avoids problems associated with the use of TEIn. However, since TMIn is solid at room temperature, it has problems such as a non-uniform evaporation rate. Knauf et al. [1988] tried to combine the advantages of TMIn (weak unwanted side reactions) with those of TEIn (liquid during use) in the new compound ethyldimethylindium (EDMIn). EDMIn is liquid at room temperature and has vapor pressure of 0.85 Torr at 17 °C, which is similar to the value of 1 Torr for TMIn and an order of magnitude greater than the vapor pressure of TEIn. However EDMIn is not as pure as the best TMIn, and TMIn remains the precursor of choice for deposition of indium-containing layers by MOCVD.

The growth of Al-containing semiconductors has always been challenging due to the fact that Al is so reactive that it readily incorporates carbon and oxygen into the solid. This is particularly problematic when TMAl is used since it pyrolyzes to form aluminum carbide. Other alkyl sources such as TEAl and TIBAl, which decompose to Al metal, can be used to reduce carbon incorporation. However, they have low vapor pressure at room temperature which is a disadvantage for MOCVD, since heating the sources above room temperature necessitates heating the gas lines and reactor tube as well. In addition, the compounds are not sufficiently stable to be effective sources. In order to solve the problem of the Al precursor, the development of new molecules with the appropriate properties is necessary. DMALH has an acceptable vapor pressure of 2 Torr at 25 °C and is found to pyrolyze at temperatures as low as 250 °C [Bhat et al. 1986]. Although the films grown by DMALH have little carbon contamination, they have very high background doping level (2×10^{18} cm^{-3}), which is caused by Si and S impurities in the DMALH source [Razeghi 1989].

Group V

For the III-V materials, the trihydrides (AsH_3, PH_3, NH,) are typically used, in spite of the fact that they are extremely toxic. Arsine (AsH_3) and phosphine (PH_3) have threshold limit values (TLV) of 0.05 and 0.3 ppm respectively. Thus, the utilization of arsine and phosphine requires costly and delicate equipment to protect the operators and the environment from possible contamination. Less toxic and hazardous alternative materials are desired for safety reasons.

Alternative materials are alkyl based, such as tertiary butyl arsine (TBA), and have now reached equivalent purity levels to the hydrides. However, tertiary butyl phosphine (TBP) is still not widely used due to possible oxygen contamination or affinity compared to PH_3. The alkyl substitutes are much more expensive than the hydrides, but are also much less toxic, with some other advantages as well. Increased production volumes will decrease the cost, but many users will not switch while the cost is high. Antimony-based compounds are usually grown from an alkyl based material like TMSb, since the Sb hydrides are very unstable. [Razeghi 1989]

1.5.3. Growth chamber designs

Design of the growth chamber is one of the most important areas in the development of MOCVD. The original research reactors fell into two main categories: either vertical reactor or horizontal reactor. These reactor designs are shown schematically in Fig. 1.10. The substrate is put on a graphite susceptor that is heated by either RF coupling via a coil surrounding the reactor, a resistant heater underneath the susceptor, or lamps placed underneath the susceptor.

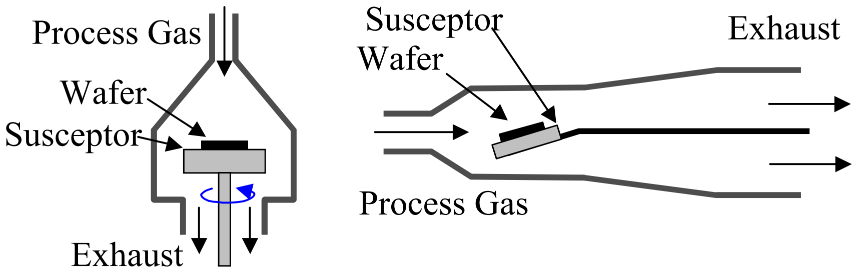


Fig. 1.10. left) Vertical and right) horizontal MOCVD reactor geometries.

The reactor wall can be cooled either with water or gas to avoid deposition onto the wall. The growth can be operated either at atmospheric or low pressure. For low pressure, the reactor pressure is typically around a tenth of an atmosphere. Low pressure is used to increase the gas velocity and help to overcome the effects of free convection from the hot substrate. Another reason for using high flow velocities is to overcome the effects of depletion of the precursor concentration at the downstream end of the deposition region.

For transport limited growth, the growth rate is determined by the rate of diffusion from the free stream to the substrate. This region is called the boundary layer. Maintaining a uniform boundary layer across the wafer improves the uniformity of the growth, but it has a cost: waste of expensive precursors and gases. The proportion of the precursors that react in the region of the substrate is very low.

This becomes a bigger problem when we want to scale to multi-wafer production. These problems have been resolved with different solutions in the vertical or horizontal reactor configurations with Emcore (now Veeco) turbo disc reactor, Aixtron closed shower head and planetary reactor, and the EMF Ltd vector flow epitaxy (VFE). All these different reactor designs have one thing in common, which is rotation of the substrate, so that the concentration of the precursors do not need to be uniform across the substrate because a portion of the substrate alternately experiences high and low concentrations that will average out. Each of these reactor designs is shown schematically in Fig. 1.11.

The turbo-disc reactor shown in Fig. 1.11a, is a vertical configuration, but the boundary layer is kept to narrow region above the susceptor by high speed rotation that pumps the gas radially outwards due to viscous drag. The rotation speeds are up to 2000 RPM in order to create this lateral flow of the constituents above the substrate. The precursors are continuously replenished from the slower downward gas stream, resulting in excellent uniformity of deposition across the wafers and a high utilization of the reactant gases. The reactor pressure is typically around 100 mBar. [Kasap et al. 2007]

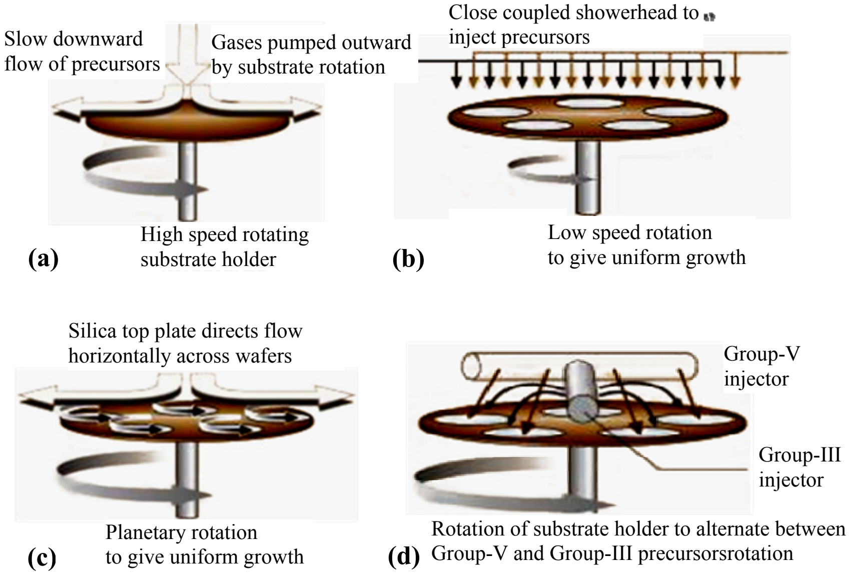

Slow downward flow of precursors
Gases pumped outward by substrate rotation
High speed rotating substrate holder
(a)
Close coupled showerhead to inject precursors
Low speed rotation to give uniform growth
(b)
Silica top plate directs flow horizontally across wafers
Planetary rotation to give uniform growth
(c)
Group-V injector
Group-III injector
Rotation of substrate holder to alternate between Group-V and Group-III precursorsrotation
(d)

Fig. 1.11. Schematic of four different production reactor designs: a) The Emcore (Veeco) Turbo-disc, b) The Aixtron closed loop showerhead, c) The Aixtron Planetary, and d) The EMF vector flow reactor.[Reproduced with permission from Chapter 14 'Epitaxial Crystal growth: Methods and Material' in 'Springer Handbook of Electronic and Photonic Materials,' S. Kasap & P. Cooper Ed. Fig. 14.14, page 285, Copyright 2006 Springer Science+Business Media Inc.]

The Close Coupled Showerhead (CCS) technology is another vertical reactor arrangement where reactants are introduced into the reactor through a water-cooled showerhead surface over the entire area of deposition. The showerhead is close to the substrate and is designed to enable precursors to be separated right up to the point where they are injected onto the substrates. The precursors are injected into the reactor chamber through separate orifices in the showerhead in order to create a very uniform distribution. Substrates are placed on top of a rotating susceptor, which is resistively heated. The three-zone heater enables adjustment of the temperature profile to provide temperature uniformity over the susceptor diameter. The susceptors are rotated at typically much lower speeds compared to the turbo-disc reactors.

The planetary reactor is horizontal flow arrangement where the reactants enter at the center of the rotation of the susceptor and flow outwards. This is an example of a fully developed flow where depletion of the reactants is occurring as the gases move away from the center. This is accentuated by a decrease in the mean flow velocity as the gases move outwards. This would normally give very poor uniformity, but the planetary rotation mechanism will rotate each wafer on the platen so they will sample alternately high and low concentrations, which gives uniform deposition. This approach has the advantages of high utilization of the precursors and the ability to extend the design to very large reaction chambers for multiple wafers. Up to 60 2-inch wafers can be held in these reactors [Kasap et al. 2007].

The fourth approach to multiple wafer deposition is the EMF Ltd vector flow epitaxy (VFE). The VFE technique utilizes individual injection of the group III and group V precursors to minimize pre-reactions and adduct formation, and the precursors can also be introduced via different carrier gases. The special injectors direct the precursors only over the wafer, thereby increasing the usage efficiency. Each injector is designed to be easily modified to allow the system to be changed freely between alloys e.g., GaN to GaAs or ZnO to InAs. The spinning wafers pass under the group III injector, which is tuned to deliver a unit mass of group III over a unit area of the wafer. Each pass under the injectors grows several uniform atomic layers. The reaction chamber can also be operated at atmospheric pressure, which simplifies the operation of the system.

1.5.4. In situ characterization

In situ characterization, which is the ability to directly observe the growth of the semiconductor material in the reactor chamber, is heavily utilized for quality control. In this section, we take a look at different *in situ* methods developed for MOCVD that deliver valuable information on the growth of the layers which speeds up the optimization loops in the development of MOCVD processes.

Reflection Anisotropy Spectroscopy (RAS)

Reflection anisotropy spectroscopy (RAS) measures the difference between the normal-incidence optical reflectance of light polarized along the two principal axes of the wafer surface as a function of photon energy (Fig. 1.12). In that sense, an RAS instrument can be considered a normal incidence ellipsometer. Since many semiconductors are cubic and therefore optically isotropic, only the anisotropy of the uppermost atomic layers will result in a change of polarization. Therefore, the method is sensitive to the properties of the wafer's growth front and can give valuable information about the doping concentrations, the composition, and the crystalline quality of the material. Fig. 1.13 shows a false color plot of the real time RAS signal for a GaAs/GaInP HBT run. The different layers are clearly identifiable in the time resolved false color plot. [Juergensen et al. 2001] This type of data, along with a database, can be used to speed up the optimization process for device fabrication.

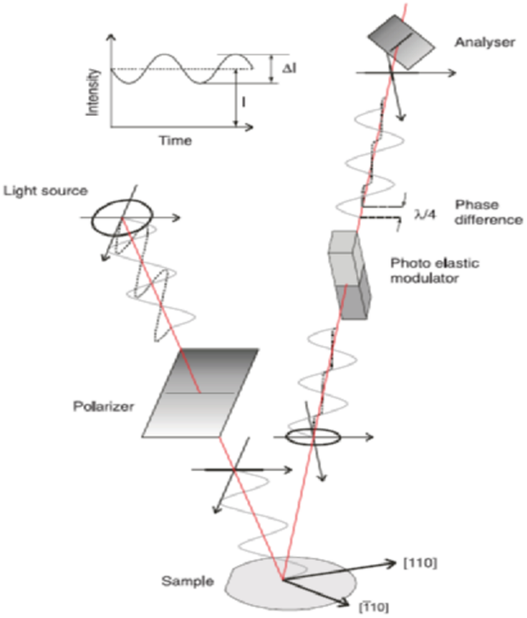


Fig. 1.12. Schematic of the principle of an RAS measurement. [Reprinted with permission from Materials Science in Semiconductor Processing Vol. 4, Juergensen, H., "MOCVD technology in research, development and mass production," fig. 8, pg. 472, Copyright 2002, Elsevier Science Ltd.]

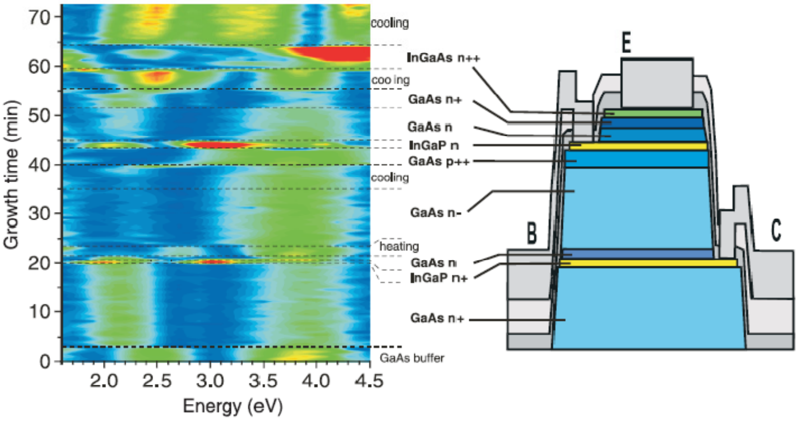


Fig. 1.13. In-situ monitoring by EpiRAS® (RAS system registered trademark of Aixtron) in the example of a HBT layer structure. [Reprinted with permission from Materials Science in Semiconductor Processing Vol. 4, Juergensen, H., "MOCVD technology in research, development and mass production," fig. 9, pg. 473, Copyright 2002, Elsevier Science Ltd.]

Reflectometer

In the case of a material system with clearly different indices of refraction between an underlying layer and the layer to be probed, time-resolved Fabry-Perot like reflectance is utilized to determine the growth-rate and the crystalline quality of the growing wafers. Besides monitoring step changes in refractive index, as a layer grows, oscillations occur due to constructive and destructive interference between reflections from the bottom and top interface of that layer. As an example, *EpiTune® I* and *EpiTune® II* are the systems used by Aixtron. [Juergensen et al. 2001] Fig. 1.14 exhibits such a reflectance trace for the case of an InGaN multi quantum well (MQW) structure. The different steps like nucleation, annealing, bulk layer and MQW growth can be clearly distinguished.

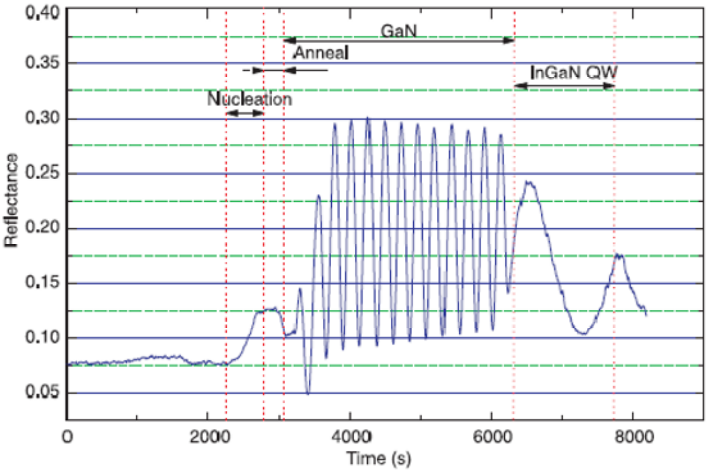


Fig. 1.14. In-situ reflectometry for an InGaN MQW growth run. [Reprinted with permission from Materials Science in Semiconductor Processing Vol. 4, Juergensen, H., "MOCVD technology in research, development and mass production," fig. 10, pg. 473, Copyright 2002, Elsevier Science Ltd.]

Emissivity Compensated Pyrometry

Measuring the real temperature of a sample being grown in MOCVD can be a challenge. A thermocouple measures the temperature in the vicinity of the sample, but not on the sample itself. A conventional pyrometer (wafer #1 in Fig. 1.15) on the other hand, looks at the sample, but it assumes that the emissivity is always constant during the growth and calculates the temperature based on that assumption. The concept of emissivity-compensated pyrometry is to measure the target thermal emission by conventional pyrometry, measure reflectivity by a reflectometer, recalculate emissivity and finally calculate target temperature by the inverse Plank formula. For non-transparent targets, energy conservation and Kirchhoff's law gives the following simple relationship between the total reflectivity (R_{total}) and emissivity:

Eq. (1.6) $\varepsilon = 1 - R_{total}$

This relation holds in the case of a specular target.

Specular opaque substrates (Si, GaAs, InP etc) (Wafer #3 in Fig. 1.15) are practically an ideal case for emissivity compensated pyrometry, which allows measuring the actual temperature of the growing layer. As long as surface is specular (as is typically for epitaxial processes), the emissivity of wafer can be measured with high accuracy using: $\varepsilon = 1 - R_{wafer.}$ Therefore any variation of emissivity due to layer growth can be compensated.

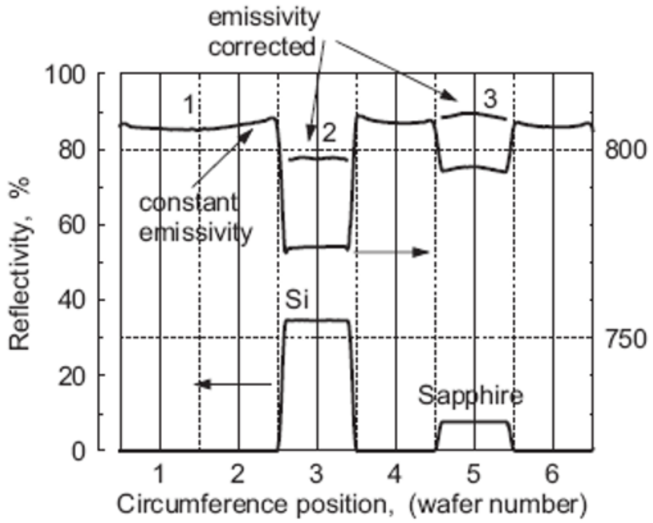


Fig. 1.15. Reflectivity and temperature profile for one wafer carrier revolution with one opaque (Si) and one transparent (sapphire) wafer and 5 empty pockets. Temperature is calculated from emission signal with emissivity: (1) preset as ε=0.85, (2) compensated for opaque wafers, and (3) compensated for transparent wafers. [Reprinted with permission from Journal of Crystal Growth Vol. 272, Belousov, M., Volf, B., Ramer, J.C., Armour, E.A., and Gurary, A., "In situ metrology advances in MOCVD growth of GaN-based materials," fig. 1, pg. 95, Copyright 2004, Elsevier B.V.]

A transparent substrate (sapphire, SiC), presented in Fig. 1.15 with a sapphire wafer (Wafer #5), does not absorb light within ultraviolet and mid-infrared bands, and, as a result, it does not emit any, according to Kirchhoff's radiation law. The best that can be done in this case is to measure the wafer carrier temperature directly under the wafer, since the emissivity of this area does not change during deposition or from run-to-run. The sampling target in this case is the area of the carrier covered by the transparent wafers. This composite target is non-transparent, and therefore its emissivity and total reflectivity are also related by Eq. (1.6). Since the surface of the carrier and the backside of the substrate are not specular, their reflectivity cannot be measured by the reflectometer, but we know that it is stable during the growth since they are covered. However, the reflectivity of the growing layer can be measured by the reflectometer. Finally, with the values for emissivity of the carrier, backside diffuse reflectivity, and the reflectivity of the growing surface, the sample emissivity can be estimated [Belousov et al. 2004].

In situ curvature and wafer tilt measurements

Strain plays a critical role in the performance of compound semiconductor devices. *In situ* stress monitors measure the curvature of the wafer and then convert it to stress values. As one example, Veeco has developed a single beam *in situ* deflectometer, which is a laser reflectometer that incorporates a Position Sensitive Detector (PSD) as a sensor. [Belousov et al. 2004] The PSD measures the intensity and position of a reflected laser beam, as illustrated in Fig. 1.16. The position of the reflected beam at the plane of the detector depends on the local surface tilt angle. As the wafer passes through the sampling beam, the angle profile as a function of wafer displacement is measured, and thus wafer curvature is calculated. Special digital processing algorithms allow the extraction of reflectivity and tilt angle data from PSD raw data.

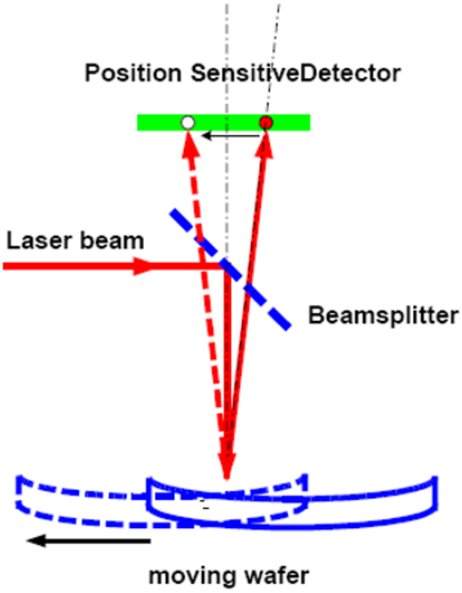


Fig. 1.16. Schematic diagram of the operation of a single beam in situ deflectometer. [Reprinted with permission from Journal of Crystal Growth Vol. 272, Belousov, M., Volf, B., Ramer, J.C., Armour, E.A., and Gurary, A., "In situ metrology advances in MOCVD growth of GaN-based materials," fig. 3, pg. 97, Copyright 2004, Elsevier B.V.]

The *in situ* deflectometer can detect changes in stress due to extremely thin layers. This is demonstrated in Fig. 1.17, where two parts of the GaN LED growth run are shown in an expanded scale. The initial nucleation layer causes the sapphire to bow, which is a result of the GaN layer strain (Fig. 1.17a). The strain can be calculated from the curvature slope and growth rate (measured from the reflectivity change). Fig. 1.17b displays the curvature versus time during growth of three very thin GaN and InGaN layers. The layer stress dependence on indium content can be estimated even for 15 Å thick layers.

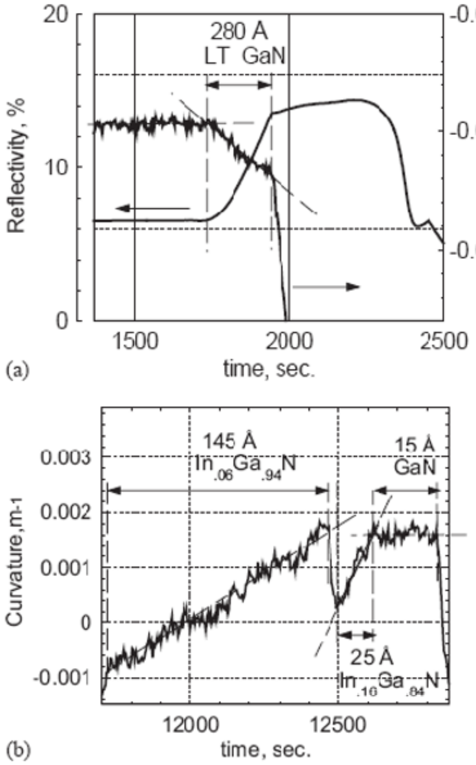


Fig. 1.17. Examples of curvature measurement. (a) Low temperature 275 Å thick GaN layer, (b) InGaN layers growth:145 Å thick $In_{0.06}Ga_{0.94}N$ (barrier), 25 Å thick $In_{0.16}Ga_{0.84}N$ (QW) and 15A° thick GaN layers. Process temperature (740 °C) is constant. The curvature jump after the $In_{0.06}Ga_{0.94}N$ layer is a result of surface temperature changes due to varying growth pressure and gas environment. [Reprinted with permission from Journal of Crystal Growth Vol. 272, Belousov, M., Volf, B., Ramer, J.C., Armour, E.A., and Gurary, A., "In situ metrology advances in MOCVD growth of GaN-based materials," fig. 5, pg. 98, Copyright 2004, Elsevier B.V.]

1.6. Molecular beam epitaxy (MBE)

1.6.1. Introduction

Molecular beam epitaxy (MBE) is an advanced technique for the growth of thin epitaxial layers of semiconductors, metals or insulators. A schematic diagram of such a system is shown in Fig. 1.18.

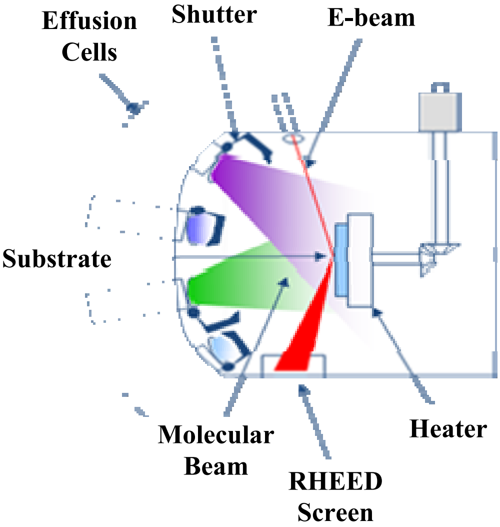


Fig. 1.18. Schematic diagram of an MBE growth chamber, showing effusion cells and shutters, the substrate stage, and the arrangement of the RHEED system.

In MBE, the sources are evaporated or sublimated in the form of beams of atoms or molecules at a controlled rate onto a crystalline substrate surface held at a suitable temperature under ultra high vacuum conditions, as illustrated in Fig. 1.18. The epitaxial layers crystallize through a reaction between the source components at the heated substrate surface. The substrate is mounted on a block at the center of the vacuum chamber and rotated continuously to promote uniform crystal growth on its surface. The thickness, composition and doping level of the epilayer can be very precisely controlled via an accurate control of the beam fluxes. The beam flux of the source materials is a function of their vapor pressure, which can be precisely controlled by their temperature. In addition, each source is mechanically shuttered, which allows for rapid modulation/ interruption of the flux at a constant source temperature and sub-monolayer control over layer thickness.

Being realized in an ultrahigh vacuum (UHV) environment, MBE can be controlled *in situ* by a large variety of surface sensitive diagnostic methods such as reflection high energy electron diffraction (RHEED), Auger electron spectroscopy (AES), x-ray photoelectron spectroscopy (XPS), low energy electron diffraction (LEED), secondary ion mass spectroscopy (SIMS) and ellipsometry. While some techniques, like ellipsometry, are applicable for MOCVD reactors, electron-based techniques can only be done in a vacuum. One of the nice features of RHEED, for example, is a direct modulation of surface reflectivity on a monolayer size scale. These powerful facilities for control and analysis eliminate much of the guesswork in MBE, and enable the fabrication of sophisticated structures using this growth technique.

Originally, molecular beam epitaxy was a UHV growth technique developed exclusively for elemental or alloy sources, where the effusion cells consisted of a resistively heated crucible. However, there are some problems associated with the use of these sources. Simple crucibles designed only for high uniformity often generate defects associated with "spitting" from the effusion cell. Also, the solid sources need to be refilled every few months. As a result, there is a long down time period necessary to reload the cells and recover the UHV condition, which increases the production costs significantly. This is even a bigger concern if multi-wafer growth is required.

In general, MBE shares many of the same advantages over LPE and VPE as MOCVD. However, compared to MOCVD, MBE growth is being done even farther from thermodynamic equilibrium. Also, as there are no organometallic sources to crack on the substrate surface, the growth rate and composition is much less sensitive to substrate temperature. Many materials can be grown at significantly lower growth temperatures in MBE, and there is even less immiscibility for alloy growth.

In the following sections, there will first be a description about diffusion cells and the different designs that have been applied to reduce spitting defects and address limited source effects. Finally, some modified versions of MBE will be presented that show some advantages over standard MBE.

1.6.2. Effusion cells used in MBE systems

Effusion cells are the basis of nearly all the beam sources in standard MBE. For different materials, different designs of diffusion are used to produce the best quality. In this section, effusion cells will be discussed along with a brief description of their operating principle and the materials they have been designed for.

Conventional Effusion Cells

In effusion cells, the solid or liquid source material is held in an inert crucible which is heated by radiation from a resistance-heated source. A schematic illustration of an effusion cell is showed in Fig. 1.19. A thermocouple is used to provide temperature control. The heater is usually a refractory metal wire either spiral wound around the crucible or is connected from end to end and is supported on insulators or inside insulating tubing. Care is taken to place the thermocouple in a position to give a realistic measurement of the cell temperature. This often takes the form of a band or spring loaded probe near the base of the crucible.

The preferred insulating material now used for sources is pyrolytic boron nitride (PBN), which can be obtained with impurity levels <10 ppm. The dissociation of this material occurs above 1400 °C. The preferred crucible material is also PBN. The standard thermocouple material employed for the sources is W-Re (5% and 26% Re). These refractory alloys are suitable for operation at high temperatures and are resistant to a reactive environment. To prevent the stainless steel outer body of the cell to get hot and start outgassing, water cooling is sometimes used.

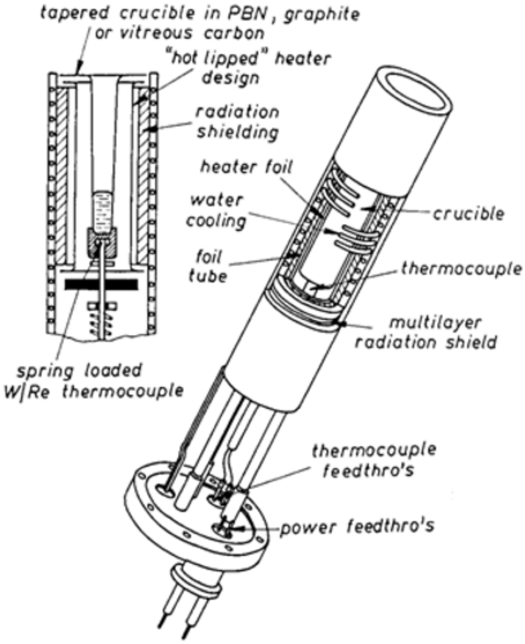

tapered crucible in PBN, graphite
or vitreous carbon
"hot lipped" heater
design
radiation
shielding
heater foil
water
cooling
crucible
foil
tube
thermocouple
multilayer
radiation shield
spring loaded
W/Re thermocouple
thermocouple
feedthro's
power feedthro's

Fig. 1.19. Diagram of an effusion cell. [Reproduced with permission from Springer-Verlag, M.A. Herman and H. Sitter, "Molecular Beam Epitaxy Fundamentals and current status," fig. 2.17, pg. 55, Copyright 1989, Springer-Verlag Berlin Heidelburg.]

SUMO Cell

One problem of the conventional effusion cells with large, tapered exit orifices is that the radiation heat loss can be significant when the beam shutter is opened. This can result in a transient temperature drop at or near the orifice. This can also result in condensation of the source material at the end of the cell and reduce the exit orifice dimensions and change the flux intensity. It can also increase the density of oval defects caused by spitting of material. Flux transients in MBE can affect interface quality and alloy composition of thin layers.

SUMO cells, which are patented by Veeco, are specifically designed to address these problems. They use two independent heater filaments for the base of the crucible and the base of the orifice. The source's thermal profile can be optimized for various materials and adjusted for the decreasing fill level over the lifetime of the charge. For most applications, the source's tip region is heated to higher temperatures than the base to compensate for radiative heat loss in the orifice region. This prevents recondensation of charge material. For high-quality Ga- and In-containing materials, this has proven also to significantly reduce Group III-related oval defects. The SUMO cell crucible has a cylindrical reservoir and a small tapered orifice which provides large capacity with good flux uniformity. You can see the shape of a SUMO cell crucible compared to a conventional cylindrical crucible in Fig. 1.20.

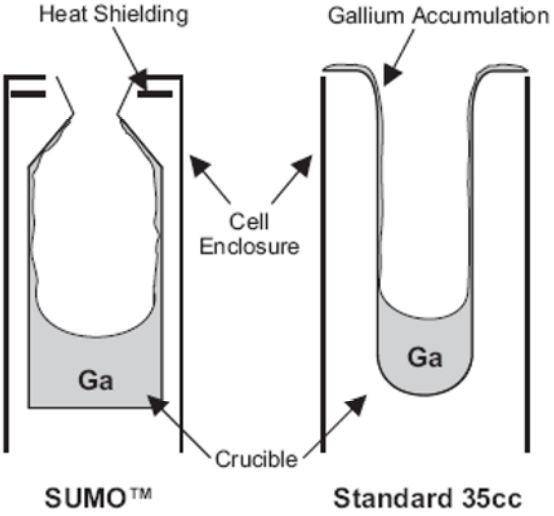


Fig. 1.20. Comparison between a sumo cell crucible and a conventional cell crucible. [Reproduced with permission from Veeco Applications Notes no. 1/98 "Improved Gallium Cell Reliability: Using the Ga SUMO Source with Ammonia in the MBE Growth of GaN," Fig. 1, pg. 2, Copyright 1998 Veeco Instruments Inc.]

Cracker Effusion Cells

Cracker cells are modified versions of conventional effusion cells with two stages. The first stage houses a standard crucible containing a solid charge and crucible. This region is heated to generate a given partial pressure of the source. In the second stage, the output of the first stage passes through a higher temperature (cracker) region which allows for homogeneous reactions of the source. As the cracker zone design typically controls the flux uniformity, this also relaxes the design constraints of the first stage reservoir. In order to minimize reactor opening for refilling purposes, the first stage reservoir is therefore typically made much larger than a standard effusion cell.

Cracker sources are often used for the group V materials (As, P, Sb), which produce tetramers (As_4, P_4, Sb_4) when directly evaporated from a standard effusion cell. In this case, the cracking region provides an elevated temperature which helps dissociate the tetramers into dimers (As_2, P_2, Sb_2). The temperatures for efficient cracking of the beams to dimers are typically within the range 800–1000 °C. [Cheung et al. 1987] Similar precautions over the choice of refractory metals and heater design are taken for the cracker region as with conventional effusion cells. It is crucial to avoid any significant heat transfer between the cracker and the cell region, otherwise uncontrolled evaporation can happen. A diagram of a cracker diffusion cell is shown in Fig. 1.21.

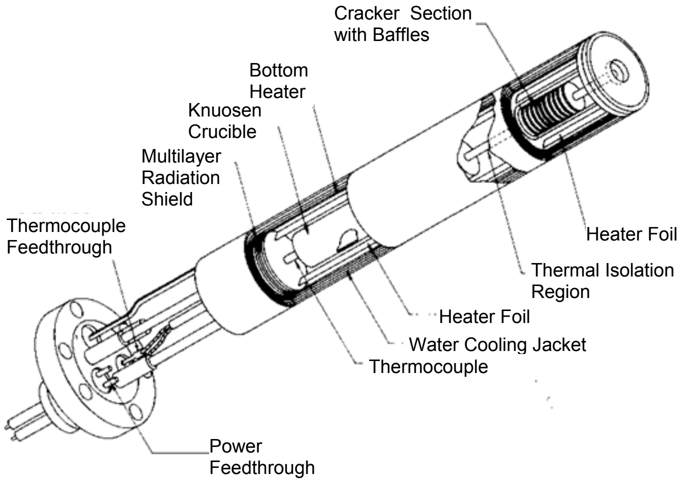

Cracker Section
with Baffles
Bottom
Heater
Knuosen
Crucible
Multilayer
Radiation
Shield
Thermocouple
Feedthrough
Heater Foil
Thermal Isolation
Region
Heater Foil
Water Cooling Jacket
Thermocouple
Power
Feedthrough

Fig. 1.21.Diagram of a cracker effusion cell. [Reproduced with permission from Parker, E.H.C., "The Technology and Physics of Molecular Beam Epitaxy, "fig. 10, pg. 29, Copyright 1985, Plenum Press, New York.]

One of the advantages of the dimer sources over tetramers is a higher sticking coefficient. For example, As_2 has been shown to have twice the sticking coefficient of As_4, which means only half of the arsenic flux should be needed per growth run. It is also known that the use of dimers provides better optical properties and lower deep level concentrations.

For large reservoir sources especially, it is difficult to change the temperature quickly and maintain stability. Modern cracker cells often incorporate another nice feature, which is a variable valve between the first and second stage. While cell flux increases exponentially with temperature, variations in flux from layer to layer in a heterostructure may only require a small adjustment to flux. As such, in order to change the flux quickly without instability, the variable valve can adjust the flux in a fairly linearly fashion without need to change the reservoir temperature. This makes a valved cracker a very useful and versatile source.

Effusion cells for II-VI compounds

The field of narrow bandgap II-VI semiconductors is dominated by $Hg_{1-x}Cd_xTe$ (MCT). The reason is that this material can cover all IR regions of interest by varying the value of x appropriately. It has also a direct band transition which allows for high absorption coefficients and quantum efficiencies of close to 100%. [Henini et al. 2002]

For MBE growth of materials with high vapor pressure, the cell construction needs to be changed. The most difficult source is Hg source, which has a vapor pressure of 2×10^{-3} Torr at 300 K. This means that the source cannot be left in the UHV chamber prior to the growth, which typically has a background pressure of 10^{-9}–10^{-10} Torr. It has also a very low sticking coefficient which means a lot of material is used in each growth. So, for MBE growth of mercury, we need a source that has the following characteristics:

- The flux should be constant throughout long growths and very reproducible from time to time.
- Since several hundred grams of Hg have to be used every day, the cell should have a large Hg capacity or should be connected to a large Hg reservoir.
- It should be possible for the mercury cell to be refilled without breaking the vacuum.

An example of a suitable mercury cell design is shown in Fig. 1.22. The Riber MCL 160 mercury cell provides a constant Hg level in a regular cell connected to a large reservoir using only gravity. The motion of the reservoir is monitored by a sensor located in the effusion cell, in order to keep the level constant. With the help of the sensor, the level is always the same. The reproducibility of such a source is excellent. The mercury temperature can also be controlled precisely, which allows for good control of the Hg flux. Lastly, since the cell itself is small, it can change temperature quickly for grading of doping or composition.

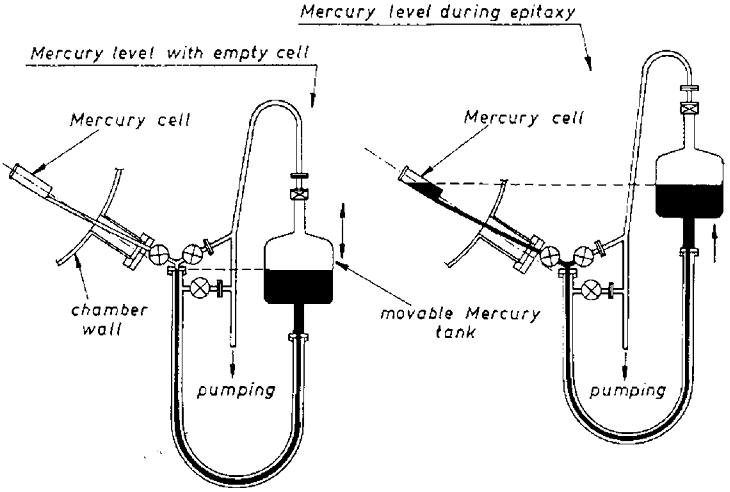


Fig. 1.22. Diagram of the operating principle of the constant level mercury source, model MCL 160 Riber [Reproduced with permission from M.A. Herman and H. Sitter, "Molecular Beam Epitaxy Fundamentals and current status," fig. 3.24, pg. 94, Copyright 1989, Springer-Verlag Berlin Heidelburg.]

1.6.3. Gas source MBE

Standard MBEs and effusion cells have limited use for multi-layer, multi-composition device structures. For abrupt compositional changes during growth using the same element, it is often necessary to have redundant sources with different temperature setpoints. Since the number of sources is fixed (typically 8–12), and sources are quite expensive, there are some practical limitations to the structure design. This is especially troublesome for quaternary materials such as $Ga_xIn_{1-x}As_yP_{1-y}$, which require multiple fluxes to be changed when changing from one composition to another.

This is typically not a problem in gas flow systems, such as MOCVD, in which the precursor flux is controlled by mass flow controllers (MFCs). In this type of system, the mass flow can be changed much more rapidly, on the order of several seconds. In addition, multiple feed lines and MFCs can branch off from the same source gas, allowing fast, direct, valve switching between various flow rates.

Gas-source molecular beam epitaxy (GSMBE) (Fig. 1.23) combines some of the best of MOCVD and MBE. This technique was first developed by Panish et al. [1980], and the name gas source MBE (GSMBE) was proposed for it. The group V effusion cells are replaced by gas injectors equipped with high temperature cracker cells. The gas injectors are valved and regulated by mass flow controllers. Typically, a hydride gas, like PH_3, is used as the source, which is cracked into various phosphorus-containing species, such as P_2. As mentioned above, the typical cracker temperature is around 800–1000 °C.

Precise on/off control of beams using fast acting gas-line valves is translated into precise control of the species arriving at the substrate. Shutters are not usually required. Atomically sharp interfaces can be achieved as a consequence of sub-monolayer switching valve times. Obviously, it can also considerably reduce the oval defect densities caused by spitting of solid sources. Further, because each injector can handle multiple gases (2–4), each controlled by MFCs, the effective number of sources available is increased dramatically.

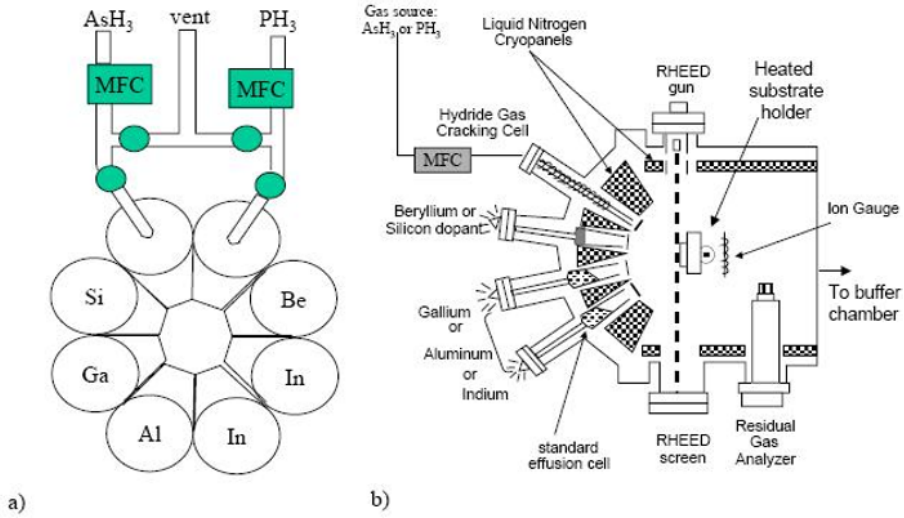

AsH3
vent
PH3
MFC
MFC
Si
Be
Ga
In
Al
In
a)
Gas source: AsH3 or PH3
Liquid Nitrogen Cryopanels
RHEED gun
Heated substrate holder
Hydride Gas Cracking Cell
MFC
Beryllium or Silicon dopant
Ion Gauge
To buffer chamber
Gallium or Aluminum or Indium
standard effusion cell
RHEED screen
Residual Gas Analyzer
b)

Fig. 1.23. EPI MOD GEN II gas-source molecular beam epitaxy reactor diagrams. a) rear view of source flange and b) side view cutaway.

Similar to MOCVD, the advantage of such a setup is an effectively unlimited supply of material (stored in gas cylinders) that can be changed or refilled without venting the growth chamber to atmosphere. The disadvantage of such a setup is the large amount of hydrogen and a small amount of toxic gas that must be pumped out of the growth chamber (AsH_3, PH_3, etc…) and post-treated before release. Even with an advanced pumping system, the growth chamber pressure rises to 10^{-5} torr during growth. However, compared to MOCVD, the gas usage is approximately 30–50 times lower for the same growth rate, because there is no laminar flow constraints in UHV. Nowadays, GSMBE remains important for the growth of InP based materials [Razeghi 2009] and also in the III-Nitride field. [Monroy et al. 2007]

1.6.4. Metalorganic MBE

A further modification of the conventional MBE is to use metalorganic sources for group III elements. This technique is called 1.6.4. Metalorganic MBE (MOMBE) or chemical beam epitaxy (CBE) [Tokumitsu et al. 1984] [Tsang et al. 1984]. Because of the low pressure in MBE, the gas transport becomes a molecular beam and the process change from the vapor deposition into beam deposition. Organometallic sources have high vapor pressure and they can produce gaseous beams without being heated significantly higher than room temperature. In short, the MOMBE reactor (Fig. 1.24) employs the gas-handling system of MOCVD and the growth chamber of a standard MBE.

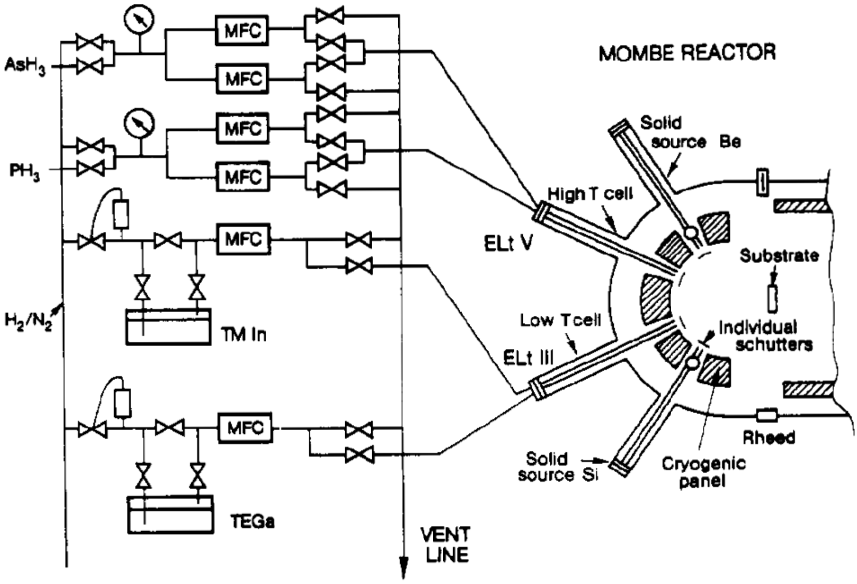

MOMBE REACTOR
AsH_3
PH_3
H_2/N_2
MFC
TM In
TEGa
VENT LINE
Solid source Be
High T cell
ELt V
Substrate
Low T cell
ELt III
Individual schutters
Rheed
Solid source Si
Cryogenic panel

Fig. 1.24. Schematic diagram of MOMBE growth chamber. [Reprinted with permission from Semiconductor Science and Technology Vol. 5, Maurel, P.H., Bove, P., Garcia, J.C., and Razeghi, M., "MOMBE growth of high quality InP and GaInAs bulk, heterojunction and quantum well layers," fig. 1, pg. 638, Copyright 1990, IOP Publishing Ltd]

In short, the advantages of the MOMBE over conventional MBE are as followed:

- The use of organometallic solid or liquid sources which do not need to be heated significantly above room temperature because of the high vapor pressure, simplifies multi wafer scale-up.
- Semi-infinite source supply, long term flux stability, and precise flow control with instant flux response
- No oval defects due to the spitting of the sources

But it has also some disadvantages which are:

- High toxicity of gas sources
- The need for gas handling and high volume pumping arrangements added to the already expensive UHV growth chamber
- No universally acceptable Al source has been found
- A very complex surface chemistry along with the temperature dependence of the surface reactions restricts the growth conditions and has a serious impact on uniformity and reproducibility of the material

At present, MOMBE has not demonstrated sufficient advantages over its parent technologies (MOCVD and MBE) to be commercially successful.

1.7. Summary

In this chapter, we have reviewed some of the current developments of semiconductor compound growth technologies. Significant technological advances have been made in response to the requirement of high material quality and larger growth scale. Over the years, many variations on these systems have been tested, but, in general, they can be classified by common characteristics which make them most suitable for specific applications. These characteristics are summarized in the following table (Table 1.2).

Growth Technique	Substrate Manu-facturing?	Crystal Purity	Alloy Miscibility Gaps?	Efficient Use of Source Material?	Hetero-structure growth?	Selective area growth	Sharp interfaces?	Costly Equip-ment?	Growth rate (μm/hr)
Czochralski/ Bridgman	Yes	Very High	Yes	Yes	No	No	No	No	Si: $>10^5$, III-V: 10^3-10^4
LPE	No	Very High	Yes	Yes	Yes	No	No	No	1-100
VPE	GaN/ AlN	Very High	Yes	No	Yes	Yes	No	Yes	1-100
MOCVD	No	High	Reduced	No	Yes	Yes	Yes	Yes	<3
MBE	No	High	Smallest	Yes	Yes	Not Standard	Yes	Yes	<3

Table 1.2. Summary of Single Crystal Growth Techniques.

References

Belousov, M., Volf, B., Ramer, J.C., Armour, E.A., and Gurary, A., "In situ metrology advances in MOCVD growth of GaN-based materials*," Journal of Crystal Growth* 272, pp. 94-99, 2004.

Bhat, R., Koza, M.A., Chang, C.C., Schwarz, S.A., Harris, T.D., *Journal of Crystal Growth* 77(1-3), pp. 7-10, 1986.

Cheung, J.T., "Role of atomic tellurium in the growth kinetics of CdTe (111) homoepitaxy," *Applied Physics Letters* 51(23), pp. 1940-1942, 1987.

Elliot, A.G., Flat, A., and Vanderwater, D.A., "Silicon incorporation in LEC growth of single crystal gallium arsenide," *Journal of Crystal Growth* 121(3), pp. 349-359, 1992.

Gao, Y.Z., Kan, H., Gao, F.S., Gong, X.Y., and Yamaguchi, T., "Improved purity of long-wavelength InAsSb epilayers grown by melt epitaxy in fused silica boats," *Journal of Crystal Growth* 234(1), pp. 85-90, 2002.

Gao, Y.Z., Gong, X.Y., Gui, Y.S., Yamaguchi, T., and Dai, N., "Electrical Properties of Melt-Epitaxy-Grown $InAs_{0.04}Sb_{0.96}$ Layers with Cutoff Wavelength of 12 µm," *Japanese Journal of Applied Physics* 43(3), pp. 1051, 2004.

Gevorkyan, V.A., "A new liquid-source version of liquid phase electroepitaxy," *Journal of Crystal Growth* 249(1-2), pp. 149-158, 2003.

Golubev, L.V., Egorov, A.V., Novikov, S.V., and Shmartsev, Y.V., "Liquid phase electroepitaxy of III-V semiconductors," *Journal of Crystal Growth* 146(1-4), pp. 277-282, 1995.

Henini, M., and Razeghi, M., *Handbook of infra-red detection technologies* Elsevier Science Ltd., 2002.

Juergensen, H., "MOCVD technology in research, development and mass production," *Materials Science in Semiconductor Processing* 4(6), pp. 467-474, 2001.

Kasap, S., and Capper, P., *Springer Handbook of Electronic and Photonic Materials*, Springer-Verlag, New York, Inc., pp. 285, 2007.

Knauf, J., "Comparison of ethyldimethylindium (EDMIn) and trimethylindium (TMIn) for GaInAs and InP growth by LP-MOVPE," *Journal of Crystal Growth* 93(1-4), pp. 43-40, 1988.

Kohiro, K., Ohta, M., and Oda, O., "Growth of long-length 3 inch diameter Fe-doped InP single crystals," *Journal of Crystal Growth* 158(3), pp. 197-204, 1996.

Meyer, M., "The Compound Semiconductor Industry in the 1990s," *Compound Semiconductors* 5, pp. 9, 1999.

Monroy, E., Guillot, F., Leconte, S., Bellet-Amalric, E., Baumann, E., Giorgetta, F.R., and Hofstetter, D., "Plasma-assisted MBE growth of nitride-based intersubband detectors," *AIP Conference Proceedings* 893(1), pp. 481-482, 2007.

Neubert, M., and Rudolph, P., "Growth of semi-insulating GaAs crystals in low temperature gradients by using the Vapour Pressure Controlled Czochralski Method (VCz)," *Progress in Crystal Growth and Characterization of Materials* 43(2-3), pp. 119-185, 2001.

Neubert, M., Rudolph, P., Frank-Rotsch, C., Czupalla, M., Trompa, K., Pietsch, M., Jurisch, M., Eichler, S., Weinert, B., and Scheffer-Czygan, M., "Crystal growth by a modified vapor pressure-controlled Czochralski (VCz) technique," *Journal of Crystal Growth* 310(7-9), pp. 2120-2125, 2008.

Pätzold, O., Wunderwald, U., Bellmann, M., Gumprich, P. Buhrig, E., Cröll, A., "New Developments in Vertical Gradient Freeze Growth," *Advanced Engineering Materials* 6(7), pp. 554-557, 2004.

Panish, M.B., "Molecular Beam Epitaxy of GaAs and InP with Gas Sources for As and P," *Journal of The Electrochemical Society* 127(12), pp. 2729-2733, 1980.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, pp. 188-193, 1989.

Razeghi, M., "High-power high-wall plug efficiency mid-infrared quantum cascade lasers based on InP/GaInAs/InAlAs material system," *Proceedings of the SPIE* 7230, p. 723011, 2009.

Rudolph, P., and Jurisch, M., "Bulk growth of GaAs An overview," *Journal of Crystal Growth* 198-199(Part 1), pp. 325-335, 1999.

Tokumitsu, E., Kudou, Y., Konagai, M., and Takahashi, K., "Molecular beam epitaxial growth of GaAs using trimethylgallium as a Ga source," *Journal of Applied Physics* 55(8), pp. 3163-3165, 1984.

Tsang, W.T., Logan, R.A., Olsson, N.A., Johnson, L.F., and Henry, C.H., "Heteroepitaxial ridge-overgrown distributed feedback laser at 1.5 μm," *Applied Physics Letters* 45(12), pp. 1272-1274, 1984.

Further reading

Kasap, S., and Capper, P., *Springer Handbook of Electronic and Photonic Materials*, Springer-Verlag, New York, Inc., 2007.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications*, Institute of Physics, Bristol, UK, 1995.

Problems

1. What is the role of the liquid encapsulation B_2O_3 in the LEC growth technique? Is the liquid encapsulation necessary in VCZ growth technique?

2. Compare the Czochralski-family and Bridgman family growth techniques.

3. Estimate the weight of a 6" in diameter and 1 m long crystal boule of Silicon.

4. Can the external reservoir configuration used for MBE growth of mercury be applied for other sources such as Arsenic, Phosphors and Antimony?

5. Why are LPE growth techniques still widely used despite the emergence of MOCVD and MBE?

2. Semiconductor Device Technology

2.1. Introduction
2.2. Oxidation
  2.2.1. Oxidation process
  2.2.2. Modeling of oxidation
  2.2.3. Factors influencing oxidation rate
  2.2.4. Oxide thickness characterization
2.3. Diffusion of dopants
  2.3.1. Diffusion process
  2.3.2. Constant-source diffusion: predeposition
  2.3.3. Limited-source diffusion: drive-in
  2.3.4. Junction formation
2.4. Ion implantation of dopants
  2.4.1. Ion generation
  2.4.2. Parameters of ion implantation
  2.4.3. Ion range distribution
2.5. Characterization of diffused and implanted layers
  2.5.1. Sheet resistivity
  2.5.2. Junction depth
  2.5.3. Impurity concentration
2.6. Summary

2.1. Introduction

In the previous Chapter, we have reviewed the various techniques used to synthesize semiconductor crystals and thin films. This represented only the first step in the fabrication of semiconductor devices. Several additional steps are necessary before a final product can be obtained, which will be described in this and the following Chapter.

In this Chapter, the discussion will be inspired from the silicon device technology because of its technological predominance and maturity in modern semiconductor industry. We will first describe and model the oxidation process used to realize a silicon oxide film. We will then discuss

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_2,
© Springer Science+Business Media, LLC 2010

the diffusion and ion implantation of dopant impurities in silicon to achieve controlled doping, and review the methods used to characterize their electrical properties. Although this Chapter discusses silicon, the methods can be equally applied to all types of semiconducting materials.

2.2. Oxidation

The ability to form a chemically stable protective layer of silicon dioxide (SiO_2) at the surface of silicon is one of the main reasons that makes silicon the most widely used semiconductor material. This silicon oxide layer is a high quality electrically insulating layer on the silicon surface, serving as a dielectric in numerous devices, that can also be a preferential masking layer in many steps during device fabrication. In this section, we will first review the experimental process of the formation of a silicon oxide. Then we will develop a mathematical model for it and determine the factors influencing the oxidation. We will then end this section by providing details on how to characterize the thickness of the formed oxide.

2.2.1. Oxidation process

A silicon dioxide layer is often thermally formed in the presence of oxygen compounds at a temperature in the range of 900 to 1300 °C. There exists two basic means of supplying the necessary oxygen into the reaction chamber. The first is in gaseous pure oxygen form (dry oxidation) through the reaction: $Si + O_2 \rightarrow SiO_2$. The second is in the form of water vapor (wet oxidation) through the reaction: $Si + 2H_2O \rightarrow SiO_2 + 2H_2$. For both means of oxidation, the high temperature allows the oxygen to diffuse easily through the silicon dioxide. The silicon is consumed as the oxide grows, and with a total oxide thickness of X, about $0.45X$ lies below the original surface of the Si wafer and $0.55X$ lies above it, as shown in Fig. 2.1. A typical oxidation growth cycle consists of dry-wet-dry oxidations, where most of the oxide is grown in the wet oxidation phase. Dry oxidation is slower and results in more dense, higher quality oxides. This type of oxidation method is used mostly for metal-oxide-semiconductor (MOS) gate oxides. Wet oxidation results in much more rapid growth and is used mostly for thicker masking layers.

Before thermal oxidation, the silicon is usually preceded by a cleaning sequence designed to remove all contaminants. Special care must be taken during this step to guarantee that the wafers do not contact any source of contamination, particularly inadvertent contact with a human person. Humans are a potential source of sodium, the element most often responsible for the failure of devices due to surface leakage. Sodium contamination can be reduced by incorporating a small percentage of chlorine into the oxidizing gas. Next, the cleaned wafers are dried and loaded into a quartz wafer holder called a boat.

The thermal oxidation process is performed with the wafers sitting in the boat loaded into a furnace where the temperature is carefully controlled. Generally, three or four separate furnaces are used in a stack manner, each with its own set of controls and quartzware. The quartz tube inside each furnace is enclosed around heating coils which are controlled by the amount of electrical current running through. A cross-section of a typical oxidation furnace is shown in Fig. 2.1.

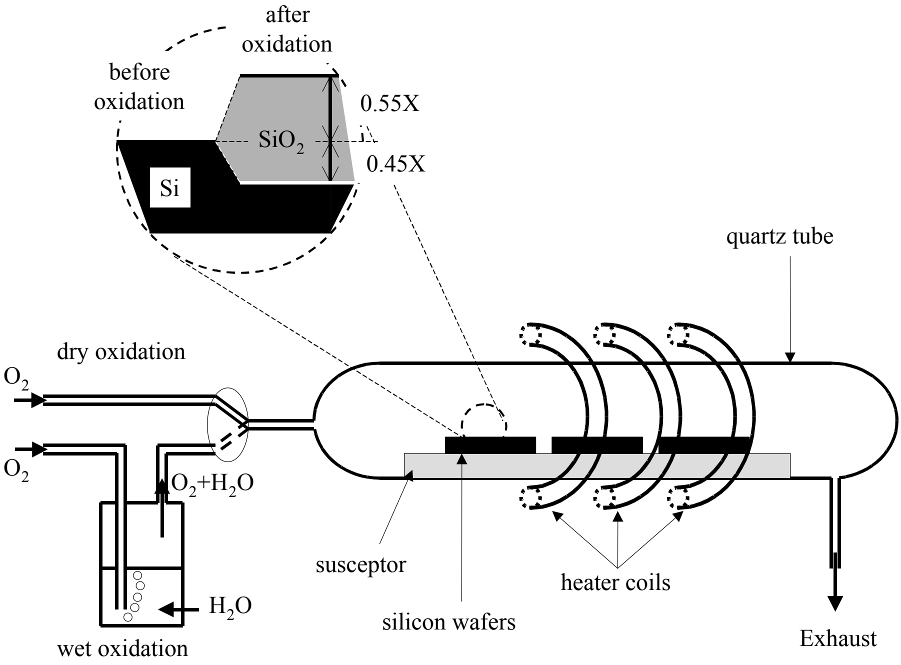


Fig. 2.1. Cross-section of an oxidation furnace: a quartz tube, heated by coils surrounding it, contains the silicon wafers in which either dry oxygen gas or water vapor can be introduced to provide the oxidizing gas. On the top left, the cross-section at the surface of a silicon wafer before and after oxidation is shown.

The furnace is suitable for either dry or wet oxidation film growth by turning a control valve. In the dry oxidation method, oxygen gas is sent into the quartz tube. High-purity gas is used to ensure that no unwanted impurities are incorporated in the layer of oxide as it forms. The oxygen gas can also be mixed with pure nitrogen gas in order to decrease the total cost of running the oxidation process, as nitrogen gas is less expensive than oxygen. In the wet oxidation method, the water vapor introduced into the furnace system is created by flowing a carrier gas into a container or bubbler filled with ultra pure water and maintained at a constant temperature below its boiling point (100 °C). The carrier gas can be either nitrogen or oxygen, and both result in equivalent oxide thickness growth rates. As the gas bubbles through the water, it becomes saturated with the water vapor. The distance to the quartz oxidation tube must be short enough to prevent condensation of the water vapor. The bubblers used in the wet oxidation process are simple and quite reproducible, but they have two disadvantages associated with the fact that they must be refilled when the water level falls too low: an improper handling of the container can result in the contamination of the water prior or during filling, and the bubbler cannot be filled during an oxidation cycle.

2.2.2. Modeling of oxidation

Using radioactive tracer experiments, the oxygen or water molecules in a dry oxidation process were found to move through the oxide film and react with the silicon atoms at the interface between the oxide film and silicon. As the oxide grows, the growth rate of the oxide layer decreases because the oxygen must pass through more oxide to reach and combine with the silicon. This is schematically illustrated in Fig. 2.2. The movement of these molecules through the forming oxide layer can be mathematically modeled using Fick's first law of diffusion:

Eq. (2.1) $$\Phi_e^{diff} = -D_n \frac{dn}{dx}$$

The objective of the following mathematical model is to determine the growth rate of the oxide layer~~,~~ that is how fast the oxide layer forms. In this model, we consider that there is a flow of a gas containing oxygen, called the oxidant, onto the sample surface, which we assume diffuses through the existing oxide layer and reacts with the underlying silicon. We will consider three different fluxes (units of particles per cm^2·s^{-1}) of oxidant, each governed by a different physical mechanism. These fluxes are shown in Fig. 2.3.

The objective of the following mathematical model is to determine the growth rate of the oxide layer~~,~~ that is how fast the oxide layer forms. In this model, we consider that there is a flow of a gas containing oxygen, called the oxidant, onto the sample surface, which we assume diffuses through the existing oxide layer and reacts with the underlying silicon. We will consider three different fluxes (units of particles per cm^2·s^{-1}) of oxidant, each governed by a different physical mechanism. These fluxes are shown in Fig. 2.3.

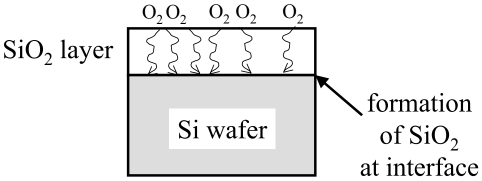

O_2 O_2 O_2 O_2 O_2
SiO_2 layer
Si wafer
formation of SiO_2 at interface

Fig. 2.2. Formation of SiO_2 in a dry oxidation process. The oxygen molecules diffuse through the existing oxide film until they reach the oxide-silicon interface where they react with silicon atoms to continue to form an oxide.

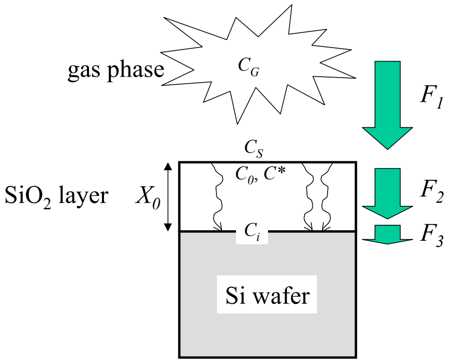


Fig. 2.3. Model for the thermal oxidation of silicon. F_1 represents the flux of oxidant from the bulk gas phase onto the sample surface, F_2 represents the flux of oxidant diffusing through the existing oxide, and F_3 represents the flux of oxidant which reaches the oxide-silicon interface and is consumed through chemical reaction with the silicon.

The first one is the flux of oxidant from the bulk gas phase onto the sample surface, denoted F_1. This flux is proportional to the difference in concentration of oxidant between the bulk gas phase and at the surface of the forming oxide:

Eq. (2.2) $$F_1 = h_G\left(C_G - C_s\right)$$

where h_G is the vapor phase mass transfer coefficient, C_G denotes the oxidant concentration in the bulk gas phase, and C_S denotes that at the surface of the forming oxide. These concentrations are generally different because some oxidant is consumed in the oxidation process. These concentrations are directly related to the partial pressures of the oxidant gas in the bulk gas phase, P_G, and at the oxide surface, P_S, through the ideal gas law:

Eq. (2.3)
$$\begin{cases} C_G = \dfrac{P_G}{k_b T} \\ C_S = \dfrac{P_S}{k_b T} \end{cases}$$

where k_b is the Boltzmann constant and T is the absolute temperature.

We can relate the oxidant concentration in the gas with the oxidant concentration in the solid phase, i.e. the oxide layer, near the surface through Henry's law:

Eq. (2.4)
$$C_0 = K_H P_S$$

where C_0 is the oxidant concentration inside the oxide layer just below its surface, K_H is Henry's law constant and P_S is the partial pressure of the oxidant in the gas phase at the oxide surface.

It will be convenient to introduce the equilibrium value of C_0, which will be denoted C^*. This concentration is related to the partial pressure in the bulk of the gas P_G through:

Eq. (2.5)
$$C^* = K_H P_G$$

Combining Eq. (2.3), Eq. (2.4) and Eq. (2.5), Eq. (2.2) can then be successively written as:

Eq. (2.6)
$$F_1 = h_G \left[\frac{P_G}{k_b T} - \frac{P_S}{k_b T} \right] = \frac{h_G}{k_b T} \left[\frac{C^*}{K_H} - \frac{C_0}{K_H} \right] = h\left[C^* - C_0\right]$$

where we have defined:

Eq. (2.7)
$$h = \frac{h_G}{k_b K_H T}$$

The second flux, denoted F_2, to consider is that of the oxidant diffusing through the oxide layer already present which can be expressed as:

Eq. (2.8)
$$F_2 = \frac{D}{X_0}\left(C_0 - C_i\right)$$

where D is the diffusion coefficient of the oxidant through the oxide, C_i is the oxidant concentration at the oxide-silicon interface, and X_0 is the thickness of the oxide.

The third flux, denoted F_3, corresponds to the incorporation of oxidant molecules which reach the oxide-silicon interface and react chemically to expand the oxide. This can be expressed as:

Eq. (2.9) $F_3 = k_s C_i$

where k_S is the chemical reaction constant for the formation of oxide.

Under steady-state conditions, these three fluxes must be equal:

Eq. (2.10) $F_1 = F_2 = F_3 = F$

This gives us three equations, for the three unknowns: C_0, C^* and C_i. Using Eq. (2.8) and Eq. (2.9) to equate F_2 and F_3, we get:

Eq. (2.11) $$C_0 = \left(1 + \frac{k_S X_0}{D}\right) C_i$$

Now, using Eq. (2.6) and Eq. (2.9) to equate F_1 and F_3, we get:

Eq. (2.12) $$C^* = C_0 + \frac{k_s}{h} C_i$$

which, after considering Eq. (2.11), becomes:

Eq. (2.13) $$C^* = \left(1 + \frac{k_S X_0}{D} + \frac{k_S}{h}\right) C_i$$

It is convenient to rearrange these relations to express C_0 and C_i as a function of C^*:

Eq. (2.14) $$C_i = \frac{1}{\left(1 + \frac{k_S X_0}{D} + \frac{k_S}{h}\right)} C^*$$

Eq. (2.15) $$C_0 = \frac{\left(1 + \frac{k_S X_0}{D}\right)}{\left(1 + \frac{k_S X_0}{D} + \frac{k_S}{h}\right)} C^*$$

We can now consider a particular case. If we assume that $h >> k_s$, i.e. the oxidation reaction at the oxide-silicon interface is much slower than the arrival of oxidant at the oxide surface, the oxidation process is then said to be interfacial reaction controlled. The Eq. (2.14) and Eq. (2.15) can then be simplified into:

Eq. (2.16) $$\begin{cases} C_i \approx \frac{1}{\left(1 + \frac{k_s X_0}{D}\right)} C^* \\ C_0 \approx C^* \end{cases}$$

Combining Eq. (2.9) and Eq. (2.16) to eliminate C_i, we can express the flux F as a function of C_0:

Eq. (2.17) $$F = \frac{k_s}{\left(1 + \frac{k_s X_0}{D}\right)} C_0$$

The rate at which the oxide layer grows is then given by the flux divided by the number N of oxidant molecules that can be incorporated into a unit volume of oxide:

Eq. (2.18) $$\frac{dX_o}{dt} = \frac{F}{N} = \frac{1}{N} \frac{k_s C_0}{\left(1 + \frac{k_s X_0}{D}\right)}$$

For dry oxidation, $N = 2.2 \times 10^{22}$ molecules per cm^3, while for wet oxidation $N = 4.4 \times 10^{22}$ molecules per cm^3. Integrating Eq. (2.18) and using the boundary condition $X_0(t=0) = X_i$, yields the following equation for X_0:

Eq. (2.19) $$X_o^2 + AX_o = B(t + \tau)$$

where τ is an integration constant and where we have denoted:

$$\text{Eq. (2.20)} \quad \begin{cases} A = \dfrac{2D}{k_s} \\ B = \dfrac{2DC_0}{N} \\ \tau = \dfrac{{X_i}^2}{B} + \dfrac{X_i}{B/A} \end{cases}$$

where X_i is the initial thickness of the oxide. For dry oxidation, an initial oxide thickness of 250 Å must be accounted for by letting $X_i = 25$ nm, in order to make Eq. (2.19) universal to both oxidation methods.

Solving for the oxide thickness in Eq. (2.19) as a function of oxidation time t, one obtains the following positive expression for X_0:

$$\text{Eq. (2.21)} \quad X_0 = \frac{A}{2}\left\{\sqrt{1 + \frac{(t+\tau)}{A^2/4B}} - 1\right\}$$

The growth time t is given directly by Eq. (2.19):

$$\text{Eq. (2.22)} \quad t = \frac{A^2}{4B}\left[\left(\frac{2X_0}{A} + 1\right)^2 - 1\right] - \tau$$

For the limiting case of "short oxidation time", where $(t+\tau) \ll A^2/4B$, we can simplify the expression in Eq. (2.21):

$$\text{Eq. (2.23)} \quad X_0 \approx \frac{A}{2}\left\{1 + \frac{1}{2}\frac{(t+\tau)}{A^2/4B} - 1\right\}$$

which is obtained after considering the Taylor expansion of the square root. We then obtain the so-called *linear oxidation law*:

$$\text{Eq. (2.24)} \quad X_0 = \frac{B}{A}(t+\tau)$$

where B/A is the linear growth rate constant, and can be calculated using Eq. (2.20) and Table 2.1.

For the other limiting case of "long oxidation time", when $t >> A^2/4B$, one obtains the *parabolic oxidation law*:

Eq. (2.25) $$X_o^2 = Bt$$

where B is the parabolic growth rate constant, and can be calculated using Eq. (2.20) and Table 2.1.

2.2.3. Factors influencing oxidation rate

Numerous factors can influence the oxidation rate by governing each of the mechanisms discussed in the previous model. For example, one of them is the diffusion coefficient in Eq. (2.8). This parameter generally follows an Arrhenius relationship as given by:

Eq. (2.26) $$D = D_0 \exp\left(-\frac{E_A}{k_b T}\right)$$

where k_b is the Boltzmann constant, E_A is the activation energy, and T is the temperature. Values for activation energy and D_0 coefficient can be found in Table 2.1. This relation indicates the strong dependence of oxide growth rate on temperature as the diffusion rate of the oxidant increases exponentially with temperature.

There exist four other factors which are commonly known to affect the oxidation rate of silicon: type of oxidation, orientation of the silicon wafer, pressure and impurity effects. For the type of oxidation, wet oxidation has a higher growth rate due to the higher solubility of the water vapor. The orientation dependence of the oxidation rate can be easily understood because the oxidation process depends on the total number of available Si atoms per unit area for oxidation at the oxide-silicon interface. Only the linear oxidation rate is expected to significantly change as a function of orientation, i.e. for short oxidation durations. For example, the oxidation rate for (111) oriented Si is faster than that for (100) oriented Si initially, in the linear region, as shown in Fig. 2.4(a) and (b). As the oxidation kinetics change from the linear rate to the parabolic rate, i.e. for longer oxidation durations, the difference between the two orientations diminishes. The pressure is proportional to the number of oxidants, and is directly proportional to both linear and parabolic growth rate constants. As can be seen in Eq. (2.19), an increase in pressure results in a slower growth rate.

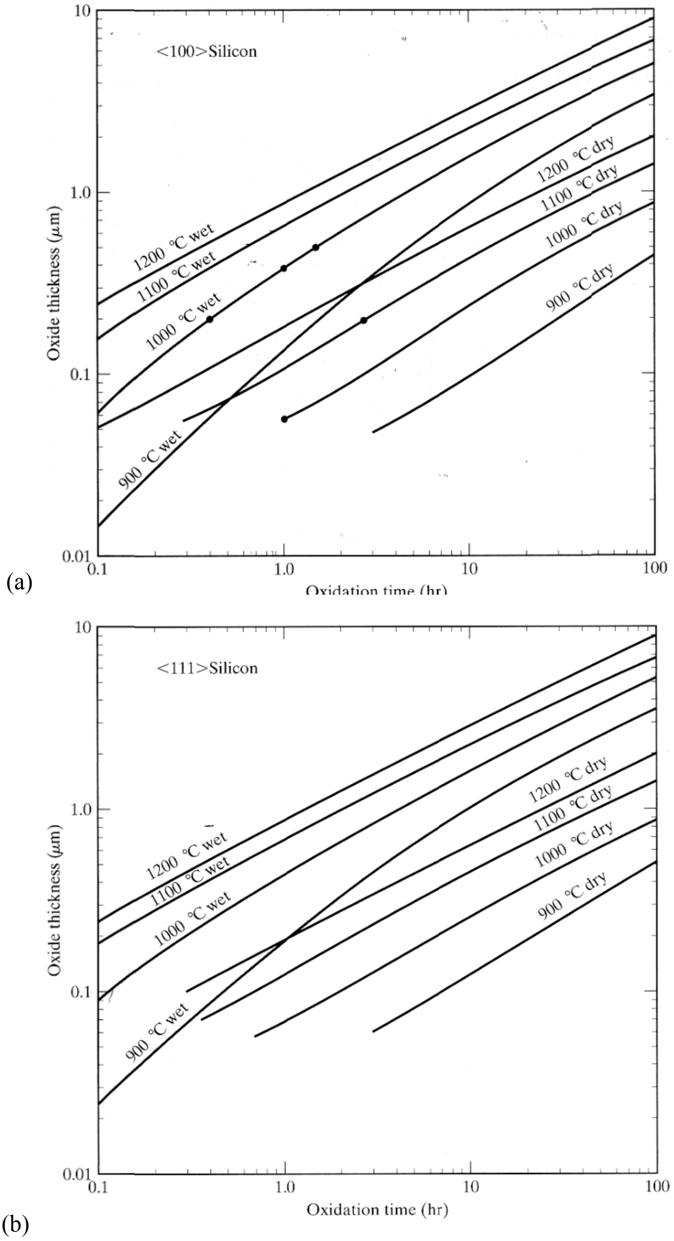


Fig. 2.4. Oxide thickness as a function of oxidation time under various conditions: (a) wet and dry oxidation of (100) silicon at several temperatures, (b) wet and dry oxidation of (111) silicon at various temperatures. [Jaeger, R.C., Introduction to Microelectronics Fabrication: Vol. 5 of Modular Series on Solidstate Devices, 2nd Edition, Fig. 3.6, p. 35.© 2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

	Wet O_2 (X_i=0 nm)		Dry O_2 (X_i=25 nm)	
	D_0	E_A (eV)	D_0	E_A (eV)
<100> Si Linear	9.7×10^7 μm·hr^{-1}	2.05	3.71×10^6 μm·hr^{-1}	2.00
<100> Si Parabolic	386 μm^2·hr^{-1}	0.78	772 μm^2·hr^{-1}	1.23
<111> Si Linear	1.63×10^8 μm·hr^{-1}	2.05	6.23×10^6 μm·hr^{-1}	2.00
<111> Si Parabolic	386 μm^2·hr^{-1}	0.78	772 μm^2·hr^{-1}	1.23

Table 2.1. D_0 coefficient values and activation energy E_A for wet and dry oxygen for different types of silicon. [Jaeger et al. 1988.]

2.2.4. Oxide thickness characterization

The accurate measurement of the thickness of a dielectric film such as silicon dioxide is very important in the fabrication of optoelectronic devices. Various techniques are available for measuring this oxide thickness, including optical interference, ellipsometry, capacitance, and the use of a color chart.

The optical interference method is a simple and nondestructive technique, which can be used to routinely measure thermal oxide thickness from less than 100 Å to more than 1 μm. The method is based on characterizing the interference pattern created by light reflected from the air/SiO_2 interface and that from the Si/SiO_2 interface, as illustrated in Fig. 2.5.

The equation governing this interference is:

Eq. (2.27) $$X_0 = \frac{\lambda(g - \Delta\varphi)}{2n^*}$$

where X_0 is the thickness of the oxide, λ is the wavelength of the incident radiation, g the order of the interference, and $\Delta\varphi$ is the net phase shift and is equal to $\varphi_s - \varphi_o$ where φ_o is the phase shift at the air/SiO_2 interface and φ_s is the phase shift at the Si/SiO_2 interface. The parameter n^* is given by:

Eq. (2.28) $$n^* = \sqrt{\overline{n}_i^2 - \sin^2\theta}$$

where $\overline{n}_i$ is the refractive index of the oxide film and θ is the angle of incidence of the light relative to the wafer. All these parameters are illustrated in Fig. 2.5.

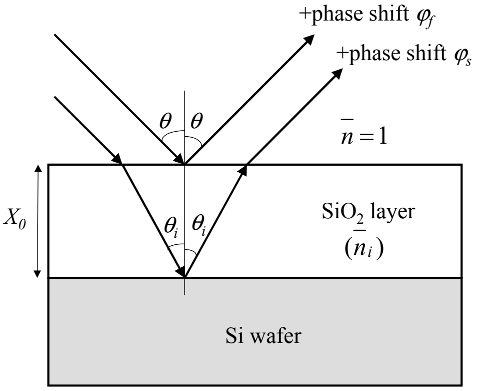


Fig. 2.5. Optical interference method for the measurement of oxide film thickness. Two rays of light with the same wavelength are shown incident on the wafer. One of them is reflected from the oxide-air interface. The other enters the oxide layer which has a different refractive index than air and is reflected at the oxide-silicon interface. A difference in optical path occurs between these two rays of light and a phase shift difference results. If the phase shift difference is an integer multiple of 2π, these two reflected rays of light interfere constructively, whereas if the phase shift difference is a half integer multiple of 2π, these rays interfere destructively.

The second method for the measurement of the oxide film thickness is ellipsometry. Ellipsometry is the most popular technique used to assess the properties of silicon dioxide films. Ellipsometry provides a non-destructive technique for accurately determining the oxide thickness, as well as the refractive index at the measuring wavelength. An illustration of an ellipsometry system is shown in Fig. 2.6. It is the most widely used tool to measure the refractive index of a wide variety of materials on any substrate, in particular SiO_2, Si_3N_4, photoresist, and aluminum oxide (Al_2O_3) on silicon substrates. Such systems can measure film thickness in the range of 20 Å to 60,000 Å with an accuracy of ±2%. An ellipsometer operates by shining polarized monochromatic light onto the wafer surface at an angle. The light is then reflected from both the oxide and the silicon surface. A phase modulation unit, numerical data acquisition and processing system work together to measure the difference in polarization. The result is then used to calculate the oxide thickness.

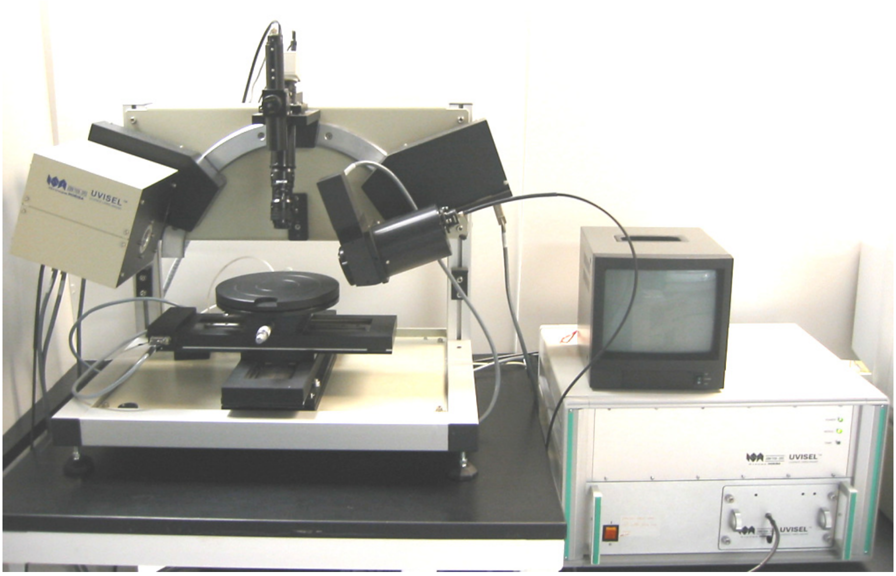


Fig. 2.6. A typical ellipsometer system, including a light source and its power supply, a sample stage, a detector and the analyzing circuits.

The third oxide film thickness measurement technique is the capacitance method, which requires the fabrication of a metal-oxide-semiconductor (MOS) capacitor. The oxide thickness is given by the following equation:

Eq. (2.29) $$X_0 = \frac{A_g \varepsilon_{ox} \varepsilon_0}{C_{ox}}$$

where C_{ox} is the experimentally measured oxide capacitance, A_g is the area of the capacitor, ε_{ox} is the dielectric constant of the oxide film, and ε_o the permittivity in vacuum.

Finally, the fourth and simplest method used to measure an oxide film thickness is by comparing the film color with a calibrated chart as shown in Table 2.2 for SiO_2. Each oxide thickness has a specific color when it is viewed under white light perpendicular to its surface. The colors are cyclically repeated for different orders of reflection.

Film thickness (µm)	Color	Film thickness (µm)	Color
0.05	Tan	0.68	Bluish
0.07	Brown	0.72	Blue green to green
0.10	Dark violet to red violet	0.77	“Yellowish”
0.12	Royal blue	0.80	Orange
0.15	Light blue to metallic blue	0.82	Salmon
0.17	Metallic to very light yellow green	0.85	Dull, light red violet
0.20	Light gold to yellow, metallic	0.86	Violet
0.22	Gold with slight yellow orange	0.87	Blue violet
0.25	Orange to melon	0.89	Blue
0.27	Red violet	0.92	Blue green
0.30	Blue to violet blue	0.95	Dull yellow green
0.31	Blue	0.97	Yellow to “yellowish”
0.32	Blue to blue green	0.99	Orange
0.34	Light green	1.00	Carnation pink
0.35	Green to yellow green	1.02	Violet red
0.36	Yellow green	1.05	Red violet
0.37	Green yellow	1.06	Violet
0.39	Yellow	1.07	Blue violet
0.41	Light orange	1.10	Green
0.42	Carnation pink	1.11	Yellow green
0.44	Violet red	1.12	Green
0.46	Red violet	1.18	Violet
0.47	Violet	1.19	Red violet
0.48	Blue violet	1.21	Violet red
0.49	Blue	1.24	Carnation pink to salmon
0.50	Blue green	1.25	Orange
0.52	Green	1.28	“Yellowish”
0.54	Yellow green	1.32	Sky blue to green blue
0.56	Green yellow	1.40	Orange
0.57	Yellow to “yellowish”	1.45	Violet
0.58	Light orange or yellow to pink borderline	1.46	Blue violet
0.60	Carnation pink	1.50	Blue
0.63	Violet red	1.54	Dull yellow green

Table 2.2. SiO_2 oxide film color chart.

Example

Q: Using the Deal-Grove oxidation model, calculate the time needed to grow a 150 nm thick oxide on top of (100) silicon by wet oxidation at a temperature of 1000 ºC.

A: T=1000 ºC=1273 K

From Table 2.1 we obtain the value of the pre-exponential factor D_0=3.71 × 10^6 µm·hr^{-1} (linear oxidation) and E_A = 2.00 eV. Using Eq. (2.26) we determine the diffusion coefficient D and then the ratio:

$$B/A = 3.71\times10^{6}\exp[-2.00/(8.617\times10^{-5})(1273)]$$

$$= 0.04478\,\mu m\cdot hr^{-1}$$

From Table 2.1 we obtain D_0 = 772 µm^2·hr^{-1} (parabolic oxidation) and E_A = 1.23 eV. Using Eq. (2.26) we calculate D and then the value for B:

$$B = 772\exp[-1.23/(8.617\times10^{-5})(1273)]$$

$$= 0.01042\,\mu m^{2}\cdot hr^{-1}$$

We can find the necessary oxidation time by using Eq. (2.19), which when rearranged becomes:

$$t = \frac{X_0{}^{2}}{B} + \frac{X_0}{B/A} - \tau$$

Since we are using wet oxidation, X_i = 0 so τ = 0,

$$\Rightarrow \quad t = \frac{(0.15)^2}{0.04478} + \frac{0.15}{0.10420} = 5.5\,.$$

Therefore, 5.5 hours is needed to grow a 150 nm thick oxide layer on (100) silicon using wet oxidation at 1000 ºC.

2.3. Diffusion of dopants

Doping is a method to control the electrical properties in semiconductors. Doping is achieved by replacing the constituting atoms of the semiconductor with atoms which contain fewer or more electrons. Through doping, the crystal composition is thus slightly altered so that it contains either a higher concentration of electrons or holes, which makes the semiconductor *n*-type or *p*-type, respectively.

The doping of semiconductors can be performed during the bulk crystal or epitaxial film growth by introducing the dopant along with the precursor chemicals. This way, the entire crystal or film is uniformly doped with the same concentration of dopants. Another method consists of carrying out the doping after the film deposition by performing the diffusion or the implantation of dopants. These have the advantage that the doping can be localized to certain regions only, by using an adequate mask to prevent the doping in undesired areas. In this section, we will focus on the diffusion of dopants and we will illustrate our discussion with the doping of silicon.

The diffusion of dopants in compound semiconductor epitaxial films generally follows a similar model. However, the effects of diffusion doping in compound semiconductor heterostructures are subtler and have been discussed in detail in specialized texts [Razeghi 1989].

2.3.1. Diffusion process

Diffusion is the process whereby a particle moves from regions of higher concentrations to regions of lower concentrations. The process could be visualized by thinking of a drop of black ink dropped into a glass of clean water. Initially, the ink stays in a localized area, appearing as a dark region in the clean water. Gradually, some of the ink moves away from the region of high concentration, and instead of there being a dark region and a clean region, there is a graduation of colors. As time passes, the ink spreads out until it is possible to see through it. Finally, after a very long time, a steady state is reached and the ink is uniformly distributed in the water. The movement of the ink from the region of high concentration (ink drop) to the region of low concentration (the rest of the glass of water) is an illustration of the process of diffusion.

In the doping of silicon by diffusion, the silicon wafer is placed in an atmosphere containing the impurity or dopant to incorporate. Because the silicon does not initially contain the dopant in its lattice, we are in the presence of two regions with different concentrations of impurities. At high temperatures (900–1200 °C), the impurity atoms can move into the crystal and diffusion can therefore occur, as schematically illustrated in Fig. 2.7.

The wafers are loaded vertically into a quartz boat and put into a furnace similar to the furnace used for oxidation. There are three types of sources to be used for the dopant atoms: solid, liquid, and gas, as shown in Fig. 2.8.

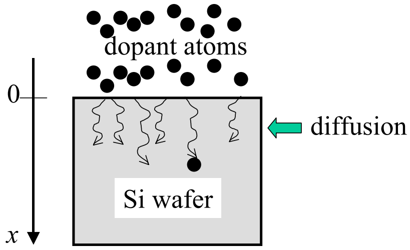

dopant atoms
0
diffusion
Si wafer
x

Fig. 2.7. Diffusion of dopants in a silicon wafer. The wafer is placed in an atmosphere containing the dopant. The gradient of dopant concentration between the atmosphere and the silicon crystal leads to their diffusion into the silicon.

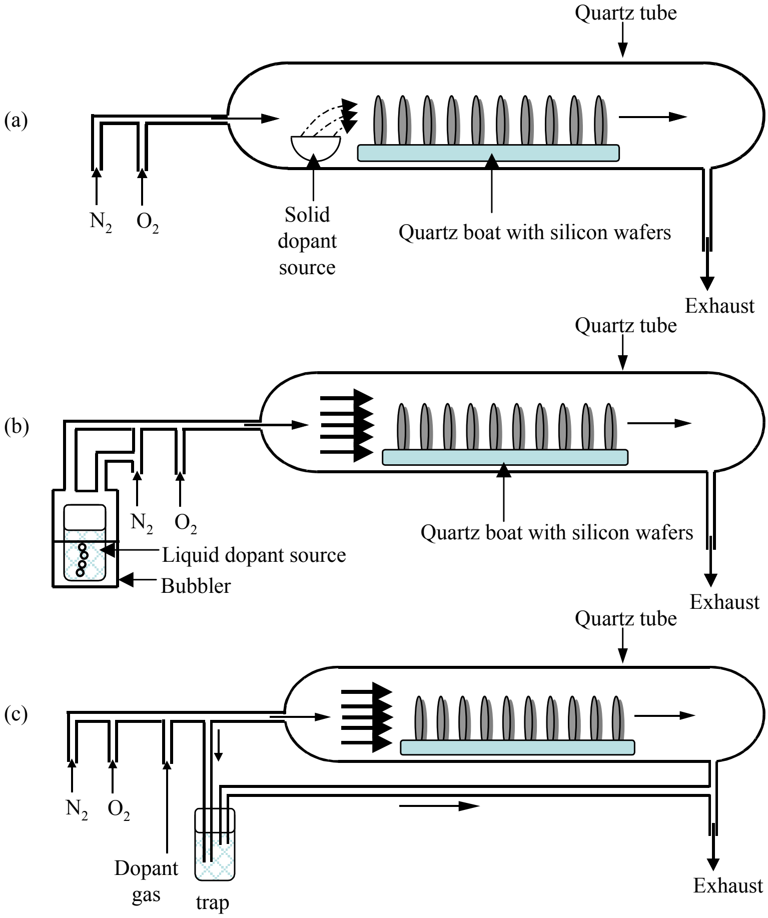
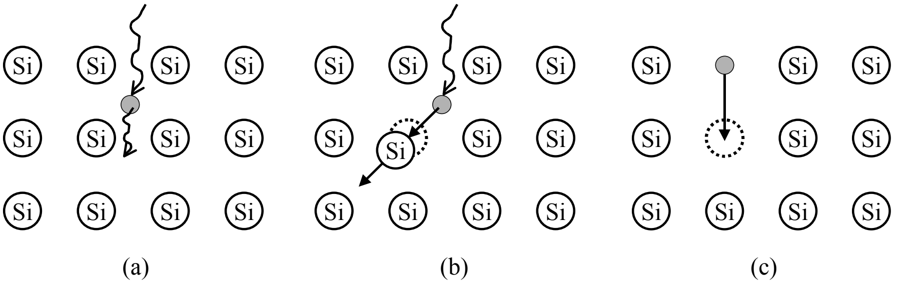

Quartz tube
(a)
N2 O2
Solid dopant source
Quartz boat with silicon wafers
Exhaust
Quartz tube
(b)
N2 O2
Liquid dopant source
Bubbler
Quartz boat with silicon wafers
Exhaust
Quartz tube
(c)
N2 O2
Dopant gas
trap
Exhaust

Fig. 2.8. Diffusion furnaces. (a) solid source diffusion with the source in a platinum source boat, (b) liquid source diffusion with the carrier gas passing through the bath, and (c) gas source diffusion with gaseous impurity sources.

There exist several types of diffusion mechanisms. An impurity can diffuse into an interstitial site in the lattice and can move from there to another interstitial site, as shown in Fig. 2.9(a). We then talk about interstitial diffusion Sometimes a silicon atom can be knocked into an interstitial site, leaving a vacancy in the lattice where a diffusing dopant atom can fit, as shown in Fig. 2.9(b). A third possible mechanism consists of a dopant directly diffusing into a lattice vacancy (Fig. 2.9 (c)). We then talk about substitutional diffusion. It is only in the cases that an impurity occupies a vacated lattice site that *n*-type or *p*-type doping can occur. The presence of such vacancies in the lattice can be due to defects or to heat which increases atomic vibrations thus giving enough energy to the silicon atoms to move out of their equilibrium positions into interstitial sites.



(a) (b) (c)

Fig. 2.9. Three possible diffusion mechanisms in a silicon wafer: (a) an impurity moves from one interstitial site to another, (b) a silicon atom is knocked into an interstitial site, thus leaving a vacancy which can be occupied by a diffusing impurity, (c) an impurity diffuses directly into a vacancy.

There are many different types of impurities that can be used for diffusion, the most common being boron, phosphorus, arsenic, and antimony. Table 2.3 lists the reactions for the materials for the three different types of diffusion sources.

The rate at which the diffusion of impurities takes place depends on how fast they are moving through the lattice. This phenomenon is quantitatively characterized by the diffusion coefficient of the impurity in silicon. Table 2.4 lists diffusion coefficient values for common impurities in silicon. We can then model the diffusion process by combining Fick's first and second law of diffusion:

Eq. (2.30) $\partial N / \partial t = D \partial^2 N / \partial x^2$

Impurity	Type	Reaction
Boron	Solid	$2(CH_3O_3)B + 9O_2 \rightarrow B_2O_3 + 6CO_2 + 9H_2O$
	Liquid	$4BBr_3 + 3O_2 \rightarrow 2B_2O_3 + 6Br_2$
	Gas	$2B_2O_3 + 3Si \rightarrow 4B + 3SiO_2$
Phosphorus	Solid	$2P_2O_5 + 5Si \rightarrow 4P + 5SiO_2$
	Liquid	$4POCl_3 + 3O_2 \rightarrow 2P_2O_5 + 6Cl_2$
	Gas	$2PH_3 + 4O_2 \rightarrow P_2O_5 + 3H_2O$
Arsenic	Solid	$2As_2O_3 + 3Si \rightarrow 3SiO_2 + 4As$
Antimony	Solid	$2Sb_2O_3 + 3Si \rightarrow 3SiO_2 + 4Sb$

Table 2.3. Diffusion reactions for common impurity types.

Element	D_0 ($cm^2 \cdot s^{-1}$)	E_A (eV)
B	10.50	3.69
Al	8.00	3.47
Ga	3.60	3.51
In	16.5	3.9
P	10.5	3.69
As	0.32	3.56
Sb	5.6	3.95

Table 2.4. Diffusion coefficient and activation energy values for common impurities in silicon.

The technology of diffusion in semiconductor processing consists of introducing a controlled amount of chosen impurities into selected regions of the semiconductor crystal. To prevent the diffusion of dopants in undesired areas, it is common to use a dielectric mask such as SiO_2 to selectively block the diffusion as show in Fig. 2.10. Fig. 2.11 shows a plot of the minimum mask thickness needed for a given diffusion time for boron and phosphorus diffusion.

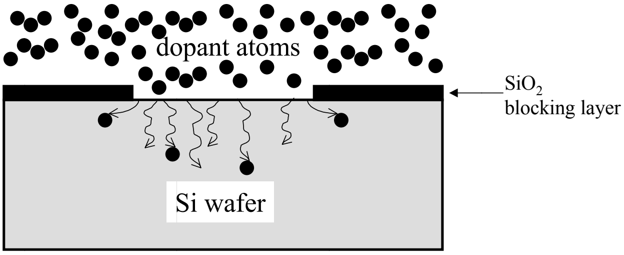


Fig. 2.10. Schematic illustration of the selective diffusion in a silicon wafer. The SiO_2 *layer acts as a blocking layer for the diffusion of dopant atoms. Some dopant atoms can diffuse laterally under the blocking layer to some extent.*

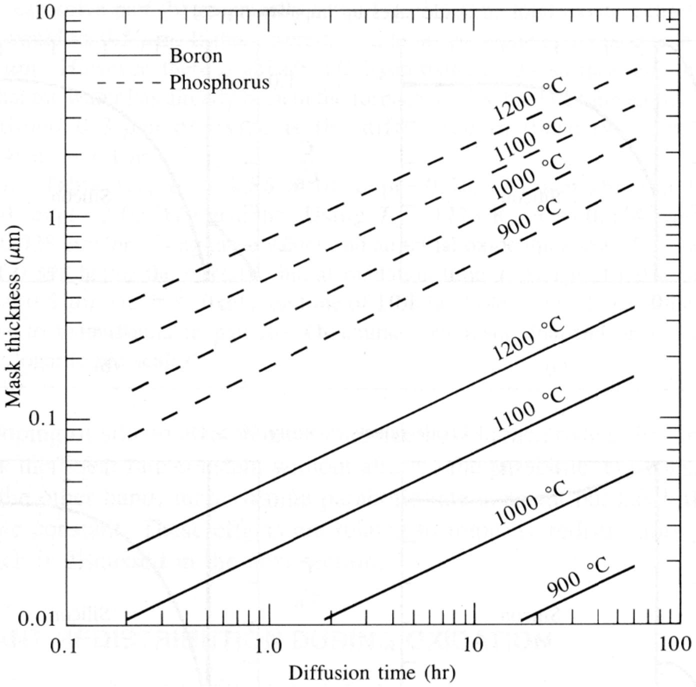


Fig. 2.11. Minimum SiO_2 *mask thickness needed for successful diffusion of boron and phosphorus in silicon for a given temperature and time. [Jaeger, R.C., Introduction to Microelectronics Fabrication: Vol. 5 of Modular Series on Solidstate Devices, 2nd Edition,© 2002, p.40, fig 3.10. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]*

There are two major techniques for conducting diffusion, depending on the state of the dopant on the surface of the wafer: (1) constant-source diffusion, also called predeposition or thermal predeposition, in which the concentration of the desired impurity at the surface of the semiconductor is kept constant; and (2) limited-source diffusion, or drive-in, in which a fixed total quantity of impurity is diffused and redistributed into the semiconductor to obtain the final profile.

2.3.2. Constant-source diffusion: predeposition

During predeposition, the silicon wafer is heated to a specific temperature, and an excess of the desired dopant is maintained above the wafer. The dopants diffuse into the crystal until their concentration near the surface is in equilibrium with the concentration in the surrounding ambient above it. At a given temperature, the maximum concentration that can be diffused into a solid is called the solid solubility. Having more dopants available outside the solid than can enter the solid guarantees that solid solubility will be maintained during the predeposition. For example, the solid solubility of phosphorus in silicon at 1000 °C is 9×10^{20} atoms/cm^3, while for boron in silicon at the same temperature it is only 2×10^{20} atoms/cm^3. These values only depend on the temperature, for a given dopant in a given semiconductor. As a result, the substrate temperature also determines the concentration of the dopant at the surface of the crystal wafer during diffusion.

Under predeposition conditions, let us denote N_0 the dopant concentration in the wafer near the surface. N_0 would be equal to the solid solubility of the dopant at the predeposition temperature if the excess dopant in the ambient above the wafer is sufficient. The concentration of dopant in the crystal at a depth x below the surface and after a diffusion time t can be known and is equal to:

Eq. (2.31) $$N(x,t) = N_0 erfc\left[\frac{x}{2\sqrt{Dt}}\right]$$

where D is the diffusion coefficient of the dopant at the predeposition temperature and *erfc* refers to the complementary error function. The complementary error function is found by complementing the integral of the normalized Gaussian function, and is shown in Fig. 2.12:

Eq. (2.32) $$erfc(\bar{x}) = 1 - erf(\bar{x}) = \frac{2}{\sqrt{\pi}}\int_{\bar{x}}^{\infty} e^{-t^2}\,dt$$

The shape of the dopant concentration function is shown in Fig. 2.13 for several values of the product Dt. We see that, as the diffusion coefficient increases or, equivalently, as the diffusion time increases, the dopant reaches deeper into the crystal. The surface concentration remains the same at N_0. The concentration N_B represents the background carrier concentration and refers to the concentration of majority carriers in the semiconductor before diffusion. The value of x for which $N(x,t)$ is equal to N_B is conventionally termed the junction depth.

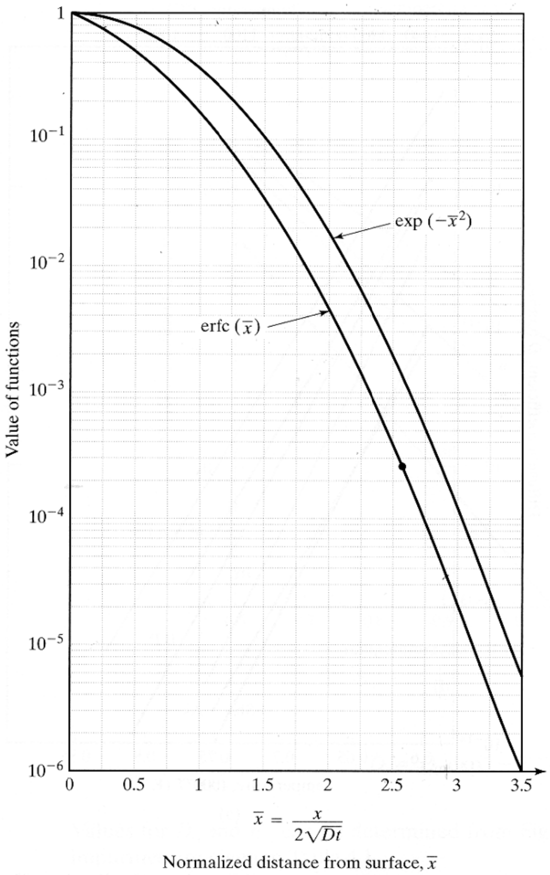


Fig. 2.12. Complementary error function, used in the calculation of the dopant concentration. [Jaeger, R.C., Introduction to Microelectronics Fabrication: Vol. 5 of Modular Series on Solidstate Devices, 2nd Edition, © 2002, p.54, fig. 4.4. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

The total amount of impurities Q introduced per unit area, also called the dose, after a diffusion duration t in a predeposition process is found by integrating the function in Eq. (2.31) for values of $x > 0$, which leads to the following expression:

Eq. (2.33) $$Q(t) = N_0\sqrt{\frac{4Dt}{\pi}}$$

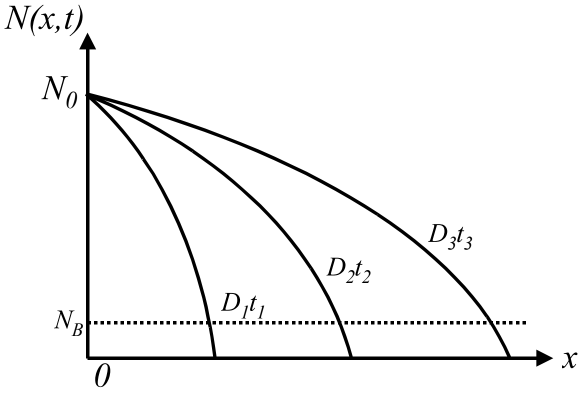


Fig. 2.13. Semi-logarithmic graph of the complementary error function, representing the dopant concentration in the crystal during predeposition, where the surface concentration is kept constant, for several values of Dt: $D_3t_3>D_2t_2>D_1t_1$. As the diffusion coefficient and/or the diffusion time is increased, the dopant reaches deeper into the crystal.

2.3.3. Limited-source diffusion: drive-in

Unlike predeposition, the drive-in diffusion process is carried out with a fixed total amount of impurity. This method allows us to better control the resulting doping profile and depth, which are important parameters in the fabrication of semiconductor devices.

During drive-in, the parameters which can be controlled include the duration of diffusion, the temperature, and the ambient gases. The dopant concentration profile of the drive-in has the shape of a Gaussian function, as shown in Fig. 2.14. In this type of diffusion, the dose remains constant causing the surface concentration to decrease. This relationship explains the shape of the curve, which can be expressed by solving Eq. (2.34) and using the boundary condition that the impurity concentration at the surface is equal to the dose:

Eq. (2.34)

$$N(x,t) = \frac{Q}{\sqrt{\pi Dt}} \exp\left[\frac{-x^2}{4Dt}\right]$$

which is expressed in units of atoms per unit volume. D is the diffusion coefficient of the impurity at the drive-in temperature and t is the drive-in time. The drive-in can be performed after a predeposition step, in a high temperature diffusion furnace once the excess dopant remaining on the surface of the wafer from the predeposition step has been removed. In this case, Q is the total dose introduced into the crystal during a predeposition step.

The limited-source diffusion process is ideally suited when a relatively low value of surface concentration is needed in conjunction with a high diffusion depth. Typically, a short period of constant-source diffusion is followed by a period of limited-source diffusion. Predeposition is used to establish the dose into a shallow layer of the surface creating a diffusion front. Then the drive-in step moves this diffusion front to the desired depth.

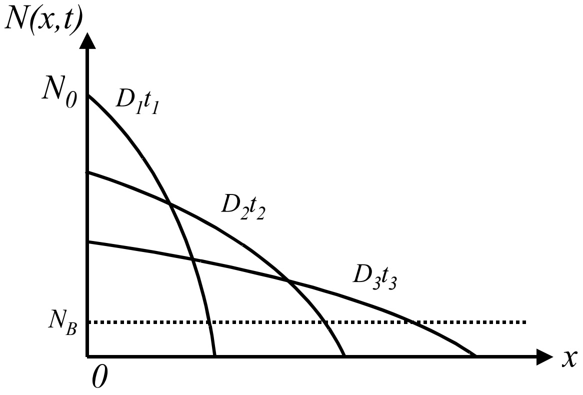


Fig. 2.14. Semi-logarithmic graph of the dopant concentration in the crystal during drive-in for several values of Dt: $D_3t_3>D_2t_2>D_1t_1$. As the diffusion coefficient and/or the diffusion time is increased, the dopant reaches deeper into the crystal. At the same time, the concentration at the surface is reduced because the drive-in is a limited-source diffusion process.

2.3.4. Junction formation

When diffusing *p*-type impurity dopants in an originally *n*-type doped semiconductor, a *p-n* junction can be formed, as shown in Fig. 2.15. In fact, the purpose of most diffusion processes is to form a *p-n* junction by changing a region of an *n*-type semiconductor into a *p*-type or vice versa.

Let us consider the example of an *n*-type doped silicon wafer which exhibits a background concentration N_B, and a *p*-type diffusing impurity with surface concentration N_0. Where the diffusing impurity profile concentration intersects the background concentration N_B, a metallurgical junction depth, x_j, is formed as shown in Fig. 2.15.

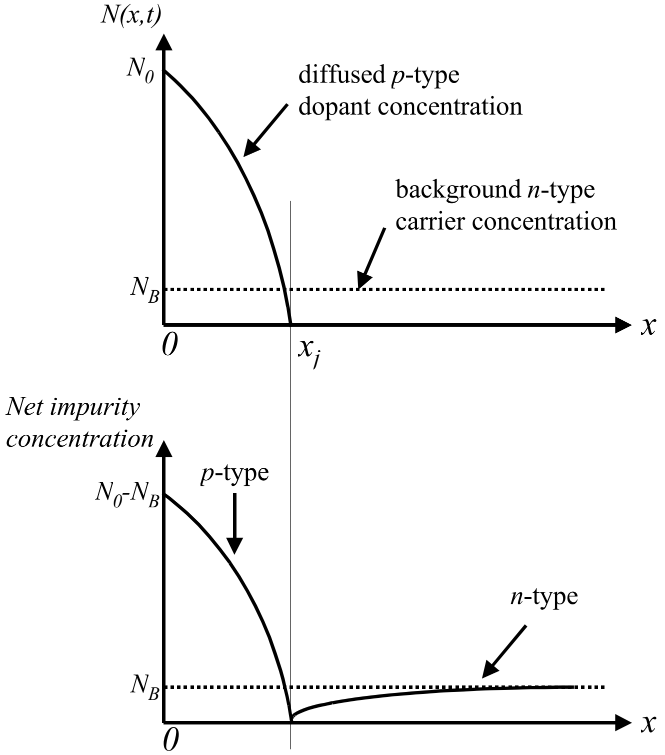


Fig. 2.15. Semi-logarithmic graphs illustrating the formation of a p-n junction through diffusion. A p-type dopant is diffused into an n-type semiconductor which has a background concentration of N_B. The p-type dopant concentration profile after diffusion is shown in the top graph. The p-n junction will occur where the p-type dopant concentration is equal to the n-type background concentration as shown on the bottom graph.

At the metallurgical junction depth, the background concentration is equal to the surface concentration, so the net impurity concentration is zero. In the predeposition process with a complementary diffusion profile, the junction depth is found by solving Eq. (2.31) and using the boundary condition that $N(x_j,t) = N_B$:

Eq. (2.35) $$x_j = \left(2\sqrt{Dt}\right) erfc^{-1}\left(\frac{N_B}{N_0}\right)$$

where $erfc^{-1}$ refers to the reciprocal function of the complementary error function. In the drive-in process with a Gaussian diffusion profile, the junction depth is found by solving Eq. (2.34) and using the boundary condition that $N(x_j,t) = N_B$:

Eq. (2.36) $x_j = 2\sqrt{Dt\ln\left(\frac{N_0}{N_B}\right)}$

By successively diffusing two impurities of different types into an originally doped wafer, more complex structures can be achieved, such as for example an *n*-*p*-*n* transistor structure as illustrated in Fig. 2.16. The starting wafer would be an *n*-type (with a background concentration N_C in this example), the first diffusion process would introduce *p*-type dopants (N_B in this example) and the second diffusion would introduce *n*-type impurities (N_E) such that $N_E >> N_B >> N_C$.

Example

Q: Calculate the dose for a boron diffusion process at 1000 ºC for 30 minutes using an *n*-type silicon substrate with a concentration of 10^{19} cm^{-3}.

A: T=1000 ºC = 1273 K
T = 30 min = 1800 s
From Table 2.4, boron has diffusion coefficient value D_0 = 10.5 cm^2·s^{-1} and activation energy E_A = 3.69 eV. Using Eq. (2.26), the diffusion coefficient becomes

$$D = 10.5\exp[-3.69/(8.617\times10^{-5}eV)(1273K)]$$

$$= 2.5822\times10^{-14}cm^2\cdot s^{-1}$$

From Eq. (2.33),

$$Q = (10^{19})\sqrt{(4\times2.5822\times10^{-14}\times1800)/\pi}$$

$$= 7.6928\times10^{13}\,cm^{-2}$$

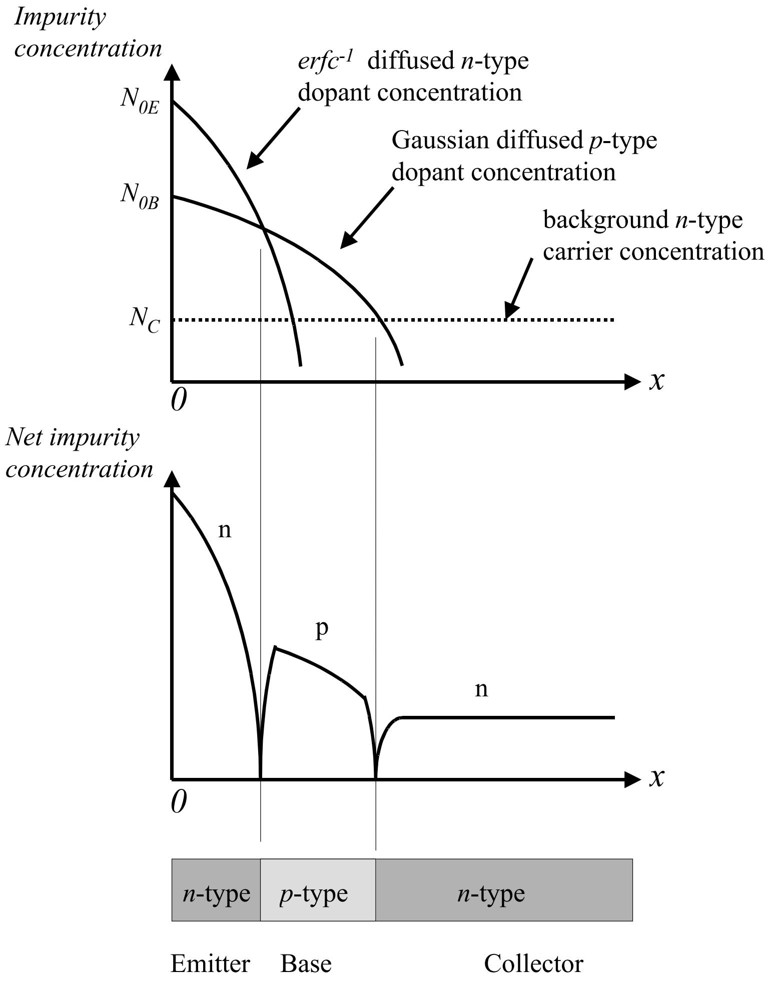

Impurity concentration
erfc^{-1} diffused n-type dopant concentration
Gaussian diffused p-type dopant concentration
background n-type carrier concentration
N_{0E}
N_{0B}
N_C
0
x
Net impurity concentration
n
p
n
0
x
n-type
p-type
n-type
Emitter
Base
Collector

Fig. 2.16. Semi-logarithmic graphs illustrating the formation of an n-p-n transistor through diffusion. p-type dopants are first diffused into an n-type semiconductor to form the first junction. n-type dopants are subsequently diffused to form the second junction.

2.4. Ion implantation of dopants

Another technique to introduce dopants into a semiconductor wafer is through ion implantation. This technique is actually a direct alternative to the thermal predeposition described previously, and can be followed by a drive-in diffusion step.

The ion implantation process selects ions of a desired dopant, accelerates them using an electric field to form a beam of ions, and scans them across a wafer to obtain a uniform predeposition dopant profile inside the crystal. The energy imparted to the dopant ion determines the ion implantation depth. Using this technique, a controlled dose of dopant impurities can be introduced deep inside the semiconductor. This is in contrast to diffusion, where the dose of dopant is introduced only at the wafer surface. In addition, like diffusion, it is possible to conduct the ion implantation in only certain well-defined areas of the wafer by using an appropriate mask. This method yields reproducible and controlled dopant concentration for semiconductor devices.

We can choose to perform selective implantation, in which regions are selectively implanted with accelerated ions by using a patterned layer of material such as silicon dioxide or photoresist, as shown in Fig. 2.17.

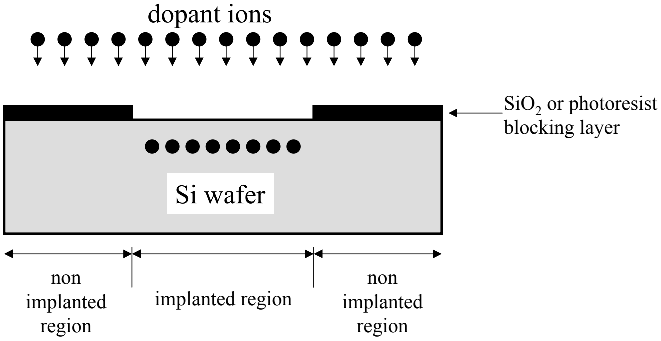


Fig. 2.17. Method of masking during ion implantation. The SiO_2 or photoresist layer acts as a blocking layer for the implantation of dopant atoms. In this process, no dopant atom can be found under the blocking layer if it is thick enough.

2.4.1. Ion generation

The first requirement of an ion implantation system is to generate ions of the desired species, accelerate them and direct them onto the wafer. A schematic of a typical ion implantation system is shown in Fig. 2.18. The dopant often comes in a gaseous form, and their ions are generated by heating them with a hot filament. These ions are then accelerated through an electric field. A magnetic field then curves the beams of ions and separates the ions, according to their atomic masses and charges, through a preset angle and output aperture. The selected ions are then further accelerated using an electric field. The beam is collimated and focused before striking the target wafer and penetrating the crystal lattice. An $x-y$ rastering mechanism ensures that a large area of the sample is scanned by the ion implantation beam.

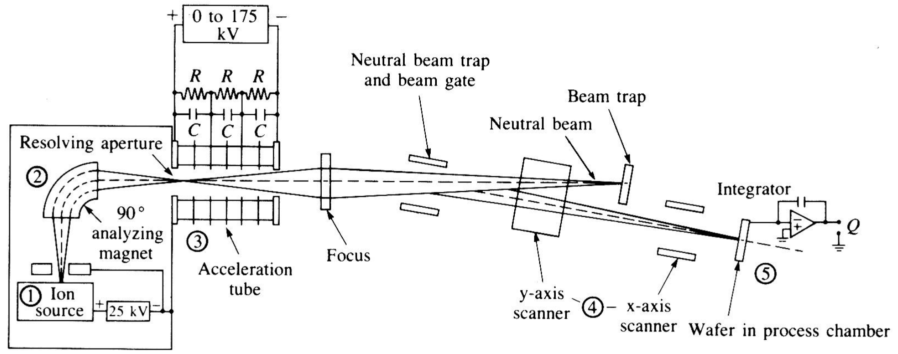

+
0 to 175 kV
−
R R R
C C C
Resolving aperture
②
90° analyzing magnet
① Ion source
+
25 kV
−
③
Acceleration tube
Focus
Neutral beam trap and beam gate
Neutral beam
Beam trap
Integrator
Q
y-axis scanner
④
x-axis scanner
⑤
Wafer in process chamber

Fig. 2.18. Schematic of a typical ion implantation equipment, including an ion source (1), a primary acceleration and an analyzing magnet (2), an acceleration tube (3), a collimating and rastering magnets (4), and the sample to be implanted. [Jaeger, R.C., Introduction to Microelectronics Fabrication: Vol. 5 of Modular Series on Solidstate Devices, 2nd Edition, Fig 5.1, p. 90., © 2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

2.4.2. Parameters of ion implantation

There are four major parameters to be controlled during ion implantation: the *energy* of the ions that reach the wafer, the dose Q of the dopant (the total number of ions that reach the wafer per unit of area), and the *depth* and *width* of the resulting implanted dopant profile.

The energy of the ions is directly controlled by the voltage used to accelerate them. It is easily understood that more energetic ions will penetrate deeper into the crystal, and potentially cause more physical damage than less energetic ones.

Because the selected ions all carry the same electrical charge, by measuring the electrical current carried by the ion beam (amount of electrical charge flowing per unit time), we can directly determine the dose. Mathematically, the latter is related to the ion beam current I through:

Eq. (2.37) $$Q = \frac{I}{qA} t$$

where q is the elementary charge, A is the implanted area and t is the duration of the ion implantation. For example, a 100 μA beam current of single ionized ions swept across a 200 cm^2 area for 60 seconds yields a dose equal to:

Eq. (2.38) $$Q = \frac{\left(100 \times 10^{-6}\right) \times (60)}{\left(1.6 \times 10^{-19}\right) \times (200)} = 1.875 \times 10^{14} \text{ dopants per cm}^2$$

By controlling the beam current and the implantation time, values of Q between 5×10^9 and 5×10^{15} cm^{-2} can typically be achieved. This range of available doses is wider than that obtainable with thermal predeposition. It is therefore possible to reach doping profiles unobtainable by any other means. If the dopants were distributed uniformly over a depth of 50 nm, the dopant concentration could be controlled between values of 10^{15} and 10^{21} dopants per cm^3.

The depth and width of the resulting implanted doping profile can be represented by the projected range and straggle, as will be discussed in the next sub-section.

2.4.3. Ion range distribution

The dopant concentration profile after implantation follows a Gaussian distribution as illustrated in Fig. 2.19. As seen in the figure, the peak concentration N_p is found at a certain depth called the projected range R_p. The projected range measures the average penetration depth of the ions.

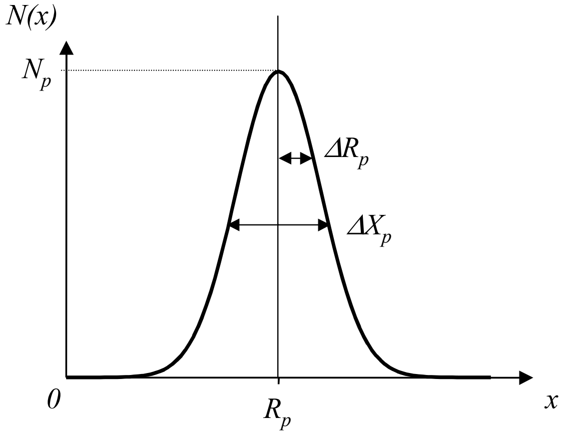


Fig. 2.19. Gaussian distribution for the concentration profile of implanted ions. The distribution is determined by its projected range, denoted R_p, corresponding to the peak concentration, and its straggle denoted by ΔR_p.

The depth at which the ions are implanted is mainly determined by the energy and the atomic number of the ions, as well as the atomic number of the substrate material. This can be easily understood because, as an impinging ion penetrates the semiconductor, it undergoes collisions with atoms and electrical repulsion with the surrounding electrons. The distance traveled between collisions and the amount of energy lost per collision are determined by a random process. Hence, even though all the ions are of the same type and have the same incident energy, they do not necessarily yield the same implantation depth. Instead, there is a distribution of depths represented by a standard deviation, called the straggle ΔR_p. Using Fig. 2.20, the impurity concentration at a given depth x can be found if the acceleration energy E_A is known:

Eq. (2.39) $$N(x) = N_p \exp\left[-\frac{(x-R_p)^2}{2\Delta R_p^2}\right]$$

For a Gaussian distribution shown in Fig. 2.19, the full width at half-maximum, denoted ΔX_p, is given by:

Eq. (2.40) $$\Delta X_p = \left(2\sqrt{2\ln 2}\right)\Delta R_p = 2.35\Delta R_p$$

The implanted dose can be determined by integrating Eq. (2.39) over all the possible depths inside the crystal:

Eq. (2.41) $$Q = \int_0^\infty N(x)dx = \sqrt{2\pi}N_p \Delta R_p$$

Example

Q: Find the full width at half-maximum for the ion-implantation using boron with an acceleration energy of 100 keV.

A: From Fig. 2.20, we obtain the normal straggle $\Delta R_p = 0.07$ µm. Using Eq. (2.40),

$$\Delta X_p = (2\sqrt{2\ln 2})(0.07) = 0.1648\,\mu m$$

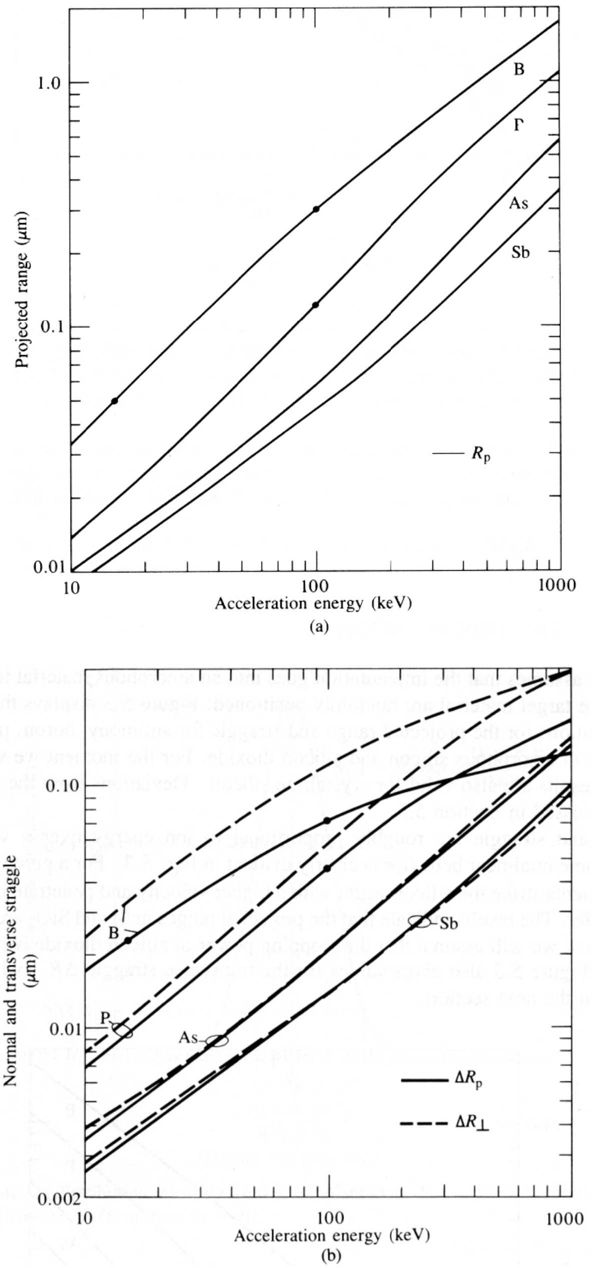


Fig. 2.20. (a) Projected range and (b) normal and transverse straggle for the ion-implantation of boron, phosphorus, arsenic, and antimony impurities for a given acceleration energy. [Jaeger, R.C., Introduction to Microelectronics Fabrication: Vol. 5 of Modular Series on Solidstate Devices, 2^{nd} Edition, © 2002, p. 93-94, fig 5.3. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

2.5. Characterization of diffused and implanted layers

Two parameters are of interest in assessing the properties of diffused or implanted layers and junctions: their electrical resistivity, junction depth, and impurity concentration. A number of techniques are available for evaluating each of these parameters.

2.5.1. Sheet resistivity

In diffused or implanted layers, we are not interested in the values of the bulk resistivity, because the carrier concentration is not uniform in space as shown in the profiles in Fig. 2.13 and Fig. 2.14. Rather, we are interested in the sheet resistivity, a quantity which can be directly measured. In order to visualize the physical meaning of this parameter, let us consider a parallelepiped semiconductor bar with a length L, a width W and a thickness H as shown in Fig. 2.21.

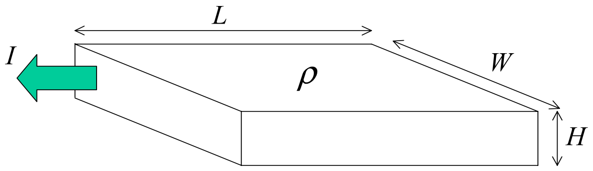


Fig. 2.21. Geometry used in determining the resistance of a block of material having uniform resistivity. When the current is flowing in the direction shown, the resistance in this direction is proportional to the length L and inversely proportional to the cross-section area WH.

From elementary physics considerations we know that the resistance of this block for a current flowing in the direction of the shown arrow is given by:

Eq. (2.42) $$R = \frac{L}{WH}\rho = \frac{L}{W}R_s$$

where ρ is the resistivity of the material, assumed to be uniform in this case. The sheet resistivity is a quantity which does not take into account the thickness of the layer, and is defined as the resistivity divided by the thickness:

Eq. (2.43) $$R_s = \frac{\rho}{H}$$

and expressed in units of “ohm per square” or Ω/□. In practice, because the thickness of the conducting layer is not always known during experiments, the sheet resistivity is the quantity that is actually measured, and the bulk resistivity is calculated subsequently.

One of the measurement methods for the sheet resistivity is the linear four-point probe method as shown in Fig. 2.22. It consists of placing four equally spaced probes on the surface of the wafer in a linear manner. The probe spacing s is typically on the order of either 1000 or 1250 µm. By sending a fixed current I through the two outer probes and measuring the voltage V across the two inner probes, we can determine the resistivity, in units of Ω·m, given by the following expression:

Eq. (2.44) $$\rho = 2\pi s\frac{V}{I} = \left[\frac{\pi t}{\ln 2}\right]\frac{V}{I}$$

Sheet resistivity can then be determined from Eq. (2.44) as:

Eq. (2.45) $$R_x = \frac{\rho}{t} = \left[\frac{\pi}{\ln 2}\right]\frac{V}{I}$$

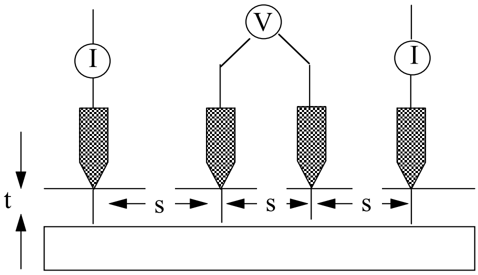


Fig. 2.22. An in-line four-point probe. Four equally spaced probes are placed on the surface of the wafer in a linear manner. A fixed current is sent through two of the outer probes and the voltage measured between the two inner probes gives a value for the sheet resistivity of the material.

Another method to measure the sheet resistivity of doped layers is van der Pauw method which can be used for any arbitrarily shaped sample of material by placing four contacts on its periphery as shown in Fig. 2.23(a). Square shaped test areas with contact regions at the four corners are usually preferred and are prepared by lithographic techniques as shown in Fig. 2.23(b).

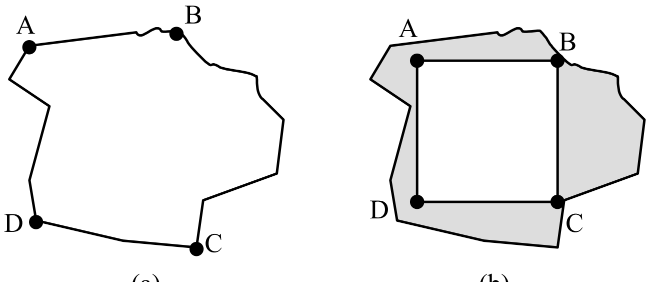


Fig. 2.23. A simple van der Pauw test structure. (a) Four contacts are placed at the periphery of any arbitrarily shaped sample and can be used to determine the resistivity of the material. (b) It is generally preferred to use a square shaped sample, which can be obtained by etching away undesired areas off the original sample.

In this configuration, a current is injected through one pair of the contacts and the voltage is measured across the remaining pair of contacts. Repeating these measurements for another pair, we are then able to define two resistances such as for example:

Eq. (2.46) $$R_{AB,CD} = \frac{V_{CD}}{I_{AB}} \text{ and } R_{BC,DA} = \frac{V_{DA}}{I_{BC}}$$

These resistances are related by the following equation relation:

Eq. (2.47) $$\exp\left[\frac{-\pi R_{AB,CD}}{R_S}\right] + \exp\left[\frac{-\pi R_{BC,DA}}{R_S}\right] = 1$$

where R_s is the sheet resistivity of the semiconductor layer. This expression allows us to implicitly determine the sheet resistivity of the sample, and thus the bulk resistivity if the layer thickness is known. For a symmetrical measurement geometry, the two resistances in Eq. (2.46) are equal and Eq. (2.47) yields a simple expression:

Eq. (2.48) $$R_s = \frac{\pi}{\ln 2} \frac{V_{CD}}{I_{AB}}$$

2.5.2. Junction depth

The junction depth x_j is defined as the distance from the top surface within the diffused or implanted layer at which the dopant concentration equals the background concentration. There exist two methods to measure the junction depth: the groove and stain method and the angle-lap method.

In the groove and stain method, a cylindrical groove is mechanically ground into the surface of the wafer as shown in the cross-section schematic in Fig. 2.24. A chemical stain creates a color contrast between the differently doped layers, thus revealing the junction.

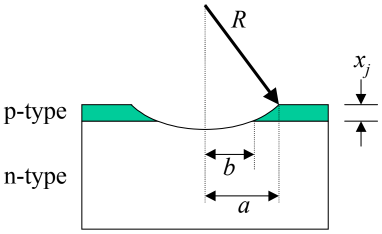


Fig. 2.24. Cross-section illustration of the junction depth measurement by the groove and stain technique. A cylindrical groove is mechanically ground into the surface of the wafer. A chemical stain creates a color contrast between the differently doped layers, thus revealing the junction.

Through purely geometrical arguments, the junction depth can be found to be equal to:

Eq. (2.49) $$x_j = \sqrt{\left(R^2 - b^2\right)} - \sqrt{\left(R^2 - a^2\right)}$$

For $R >> a$ and b, this expression can be simplified into:

Eq. (2.50) $$x_j \approx \frac{\left(a^2 - b^2\right)}{2R}$$

In the angle-lap method, a piece of the wafer is mounted on a special fixture which permits the edge of the wafer to be lapped at an angle between 1 and 5° as shown in Fig. 2.25. The sample is then stained with, for example, a 100 ml hydrofluoric acid/nitric acid solution. Once stained, the sample is observed under a collimated monochromatic light at normal incidence.

An interference pattern can be observed through a cover glass, and the junction depth may be calculated by counting the interference fringes and then applying the equation:

Eq. (2.51) $$x_j \approx d \tan\theta = \frac{N\lambda}{2}$$

where θ is the angle of the etched lap, λ is the wavelength of the monochromatic light and N is the number of fringes.

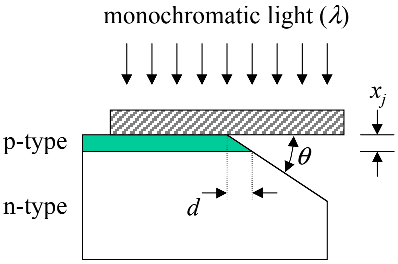


Fig. 2.25. Junction depth measured by the angle-lap method. A piece of the wafer is mounted on a special fixture which permits the edge of the wafer to be lapped at an angle. The sample is then stained and then observed under a collimated monochromatic light at normal incidence. An interference pattern can be observed through a cover glass and the junction depth can be calculated by counting the number of interference fringes.

2.5.3. *Impurity concentration*

There are many ways to measure the impurity concentration of a sample, but one of the most common techniques is Secondary Ion Mass Spectroscopy (SIMS). A schematic representation of the experimental setup is shown in Fig. 2.26. SIMS is a destructive characterization technique that operates with a highly energetic beam of ions hitting the sample, causing the sputtering or ejection of atoms from the sample material. Some of these ejected atoms are charged ions, or secondary ions. A mass spectrometer then separates and collects the secondary ions. The number of collected secondary ions then allows a detector to determine the material composition. An example plot of the impurity concentration information obtained from SIMS is shown in Fig. 2.27.

SIMS is an excellent technique for identifying all types of elements, unlike other measurement resources. One disadvantage to the SIMS measurement is the sensitivity of the technique. The sensitivity can be affected by the built-up charge from the sputtering process, type of ion beam. Another disadvantage to SIMS is the limitation of the beam area that hits the sample.

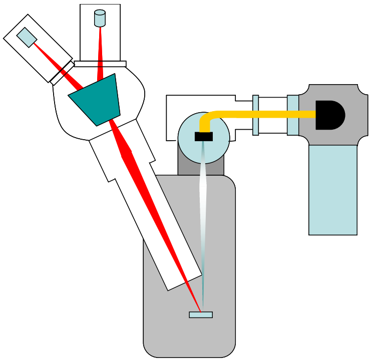


Fig. 2.26. Secondary Ion Mass Spectrometer. Low energy ion-beam hits the sample surface and atoms sputter off of it. The atoms are then ionized and sent through a mass spectrometer, where secondary ions are collected. The mass spectrometer identifies the atomic species and then the detector uses these secondary ions to determine the profile as a function of depth.

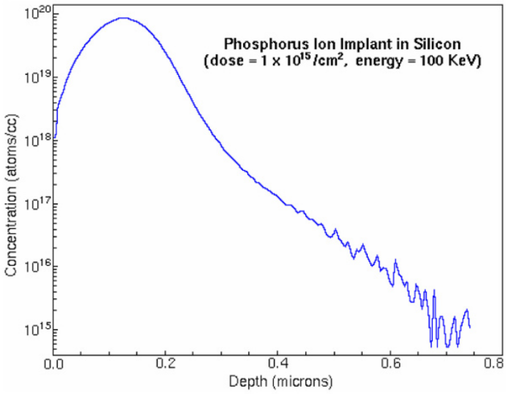


Fig. 2.27. Secondary Ion Mass Spectroscopy plot measuring the impurity concentration as a function of depth for ion implantation of phosphorus into silicon. [From http://www.me.ust.hk/~mejswu/MECH343/343SIMS.pdf.]

2.6. Summary

In this Chapter, we have reviewed a few of the steps involved in the fabrication of semiconductor devices, including oxidation, diffusion and ion implantation. Although the discussion was primarily based on silicon, the concepts introduced are applicable for the entire semiconductor industry.

We described the oxidation experimental process, mathematically modeled the formation of a silicon oxide film, discussed the factors influencing the oxidation and reviewed the methods used to characterize the oxide film. The diffusion and ion implantation of impurity dopants in silicon to achieve controlled doping in selected areas of a wafer was described, along with the resulting dopant concentration profiles inside the semiconductor. The predeposition and drive-in conditions of diffusion were discussed. Methods used to assess the electrical properties of the diffused or implanted layers were reviewed.

References

Jaeger, R.C., *Introduction to Microelectronic Fabrication*, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 2002.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, pp. 188-193, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications*, Institute of Physics, Bristol, UK, 1995.

Further reading

Campbell, S.A., *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, New York, 1996.

Diaz, J.E., *Fabrication of High Power Aluminum-Free 0.8 μm to 1.0 μm InGaP/InGaAsP/GaAs Lasers for Optical Pumping*, Ph.D. dissertation, Northwestern University, 1997.

Fogiel, M., *Microelectronics-Principle, Design Techniques, and Fabrication Processes*, Research and Education Association, New York, 1968.

Ghandi, S., *VLSI Fabrication Principles*, John Wiley & Sons, New York, 1983.

Soclof, S., *Design and Application of Analog Integrated Circuits*, Prentice-Hall, Englewood Cliffs, NJ, pp. 8-23, 1991.

Streetman, B.G., *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, pp. 65-70, 1980.

Trumbore, F.A., "Solid solubilities of impurity elements in germanium and silicon," *Bell System Technical Journal* 39, pp. 205-233, 1960.

Problems

1. A (100) Si wafer undergoes the following sequence of oxidation steps: one-hour dry oxidation at 1100 °C, two-hour wet oxidation at 1000 °C, and one-hour dry oxidation at 1100 °C. Calculate the thickness after each oxidation step.

2. Compare the thickness of silicon dioxide film grown on (100) Si for wet and dry oxidation at 1100 °C. Compare the two different orientations (100) and (111) Si using the same conditions for wet oxidation.

3. Calculate the time required to grow 2 μm silicon dioxide on (111) silicon wafer for wet oxidation at 1050 °C. How long would it take to grow an additional 5 μm?

4. A silicon oxide layer is grown for 65 minutes at 1100 °C on (111) silicon by passing oxygen through a 95 °C water bath. How thick an oxide layer is grown on the silicon surface?

5. A (100) oriented silicon wafer is already covered with a 0.5 μm thick oxide film. How long would it take to grow an additional 0.1 μm oxide using wet oxidation at 1373 K? Compare the result with the linear oxidation law using a rate of $B/A = 3\ \mu\text{m}\cdot\text{hr}^{-1}$.

6. An *npn* transistor is formed by boron diffusion on an *n*-type silicon wafer with impurity concentration of $10^{20}\ \text{cm}^{-3}$ and doping concentration $10^{16}\ \text{cm}^{-3}$. Constant source diffusion is performed for 30 minutes followed by limited-source diffusion for 2 hours, both at 1000 °C. Find the junction depth after each step.

7. What is the required thickness for a SiO_2 mask used for selective phosphorus diffusion. The diffusion was performed at 1000 °C for 3 hours.

8. An impurity is diffused into silicon in the constant-source diffusion case with a surface concentration $N_0 = 10^{19}\ \text{cm}^{-3}$. The diffusion coefficient is known to be equal to $2 \times 10^{-12}\ \text{cm}^2\cdot\text{s}^{-1}$ at 1100 °C with an activation energy of 4 eV. A diffusion length of 1 μm is aimed at. (a) After diffusion at 1000 °C, what is the total dose diffused in the layer? (b) How long must the diffusion last?

9. An impurity is diffused into silicon at 1100 °C during 20 minutes. A diffusion length of 1 μm is measured and the dose diffused in the sample is $2 \times 10^{15}\ \text{atoms}\cdot\text{cm}^{-2}$. Determine the diffusion coefficient. Determine the surface concentration, assuming constant-source diffusion.

10. Find the dose in silicon for phosphorus that is implanted with an energy of 100 keV with a 0.1 μm SiO_2 layer and peak concentration of

10^{17} cm^{-3}. Find the time required for this implantation onto a 2” wafer using 2 μA of current.

11. Implantation using phosphorus is done such that the implantation peak is located at the Si-SiO$_2$ interface with an energy of 40 keV. The dose is 5.24×10^{14} cm^{-2}, and background concentration is 3×10^{16} cm^{-3}. Find the minimum oxide thickness required for a masking layer.

12. Compare the time required for implantation of phosphorus and boron with an energy of 75 keV, with a desired peak concentration of 10^{18} cm^{-3}, into a 140 mm silicon wafer with 1 μA.

13. A constant voltage of 5 V is applied to each of the five contact pads on a given sample. The location of the pads are at $x_1 = 1$, $x_2 = 5$, $x_3 = 10$, $x_4 = 16$, and $x_5 = 23$ μm. The width is the same for each contact and is given as 1 μm. The thickness is given as 200 μm. Find the contact resistance using transmission line measurement if the current is measured as $I_{12} = 20$ mA, $I_{23} = 17$ mA, $I_{34} = 10$ mA, and $I_{45} = 3$ mA.

14. Calculate the resistivity using the van der Pauw method with measured current $I_{AB} = 1$ mA and $V_{CD} = 2$ V. The length $L = 1$ mm and width $W = 2$ mm.

15. Assume that an As implant leads to a uniform electron concentration of 10^{19} cm^{-3} down to a depth of 0.1 μm and a mobility of 100 cm^2·V^{-1}·s^{-1}. Determine the resistivity and the sheet resistivity of the implanted layer. If a square van der Pauw pattern with 1 cm side length is used with a 10 V applied at two adjacent contacts, what current would be measured through the other two contacts?

3. Semiconductor Device Processing

3.1. Introduction
3.2. Photolithography
  3.2.1. Wafer preparation
  3.2.2. Positive and negative photoresist
  3.2.3. Mask alignment and fabrication
  3.2.4. Exposure
  3.2.5. Development
  3.2.6. Direct patterning and lift-off techniques
  3.2.7. Alternative lithographic techniques
3.3. Electron-beam lithography
  3.3.1. Electron-beam lithography system
  3.3.2. Electron-beam lithography process
  3.3.3. Parameters of electron-beam lithography
  3.3.4. Multilayer resist systems
  3.3.5. Examples of structures
3.4. Etching
  3.4.1. Wet chemical etching
  3.4.2. Plasma etching
  3.4.3. Reactive ion etching
  3.4.4. Sputter etching
  3.4.5. Ion milling
3.5. Metallization
  3.5.1. Metal interconnections
  3.5.2. Vacuum evaporation
  3.5.3. Sputtering deposition
3.6. Packaging of devices
  3.6.1. Dicing
  3.6.2. Wire bonding
  3.6.3. Packaging
3.7. Summary

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_3,
© Springer Science+Business Media, LLC 2010

3.1. Introduction

The first few important steps in the process of fabricating semiconductor devices, involve oxidation, diffusion and ion implantation of dopants. The resulting semiconductor wafer then undergoes a series of additional steps before the final device is obtained, which are shown in the flowchart in Fig. 3.1. The major ones include lithography, etching, metallization and packaging, each of which will be discussed in the present Chapter.

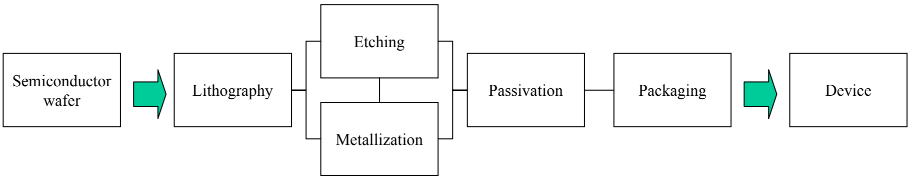


Fig. 3.1. Basic fabrication process flowchart, including the important steps: lithography, etching, metallization, passivation and packaging, which transform a semiconductor wafer into a device that can be used in electronic systems.

3.2. Photolithography

Lithography consists of preparing the surface of a semiconductor wafer in order to allow the subsequent transfer of a specific pattern. To do so, the surface of the semiconductor must be carefully prepared and a film called resist is conformally applied onto it. Parts of this resist film will be selectively "activated" through a number of processes, while others will be left untouched in order to transfer the desired pattern. This is generally achieved through what is called mask alignment and exposure. A mask is a template which contains the desired pattern to be transferred. Finally, the resist is developed to reveal the desired pattern before proceeding to the subsequent processing steps.

There are several types of lithography techniques depending on the method used to activate the resist film. The most common form of lithography uses ultraviolet light and is called photolithography. This is currently the most widely used technique in microelectronics industry and is routinely employed to achieve features as small as 0.18 µm. In this section, we will describe in detail the photolithography method.

3.2.1. Wafer preparation

We are assuming that we have completed the basic steps of making the film and that we know how to grow barrier and oxide layers. Now we can focus on the process of creating a pattern on the surface of the semiconductor wafer can now be started.

The most important step in the lithographic process is the minimization of defects caused by particles either falling on the surface prior to epitaxy or during the steps of lithography. This is because even the smallest dust particulate on a chip would destroy dozens if not hundreds of transistors on a state of the art microprocessor. This is the chief reason why semiconductor processing is performed in ultra clean facilities, know as a cleanrooms. A cleanroom uses sophisticated filtration to remove airborne particulates and are rated by the maximum number of particles per volume of air.

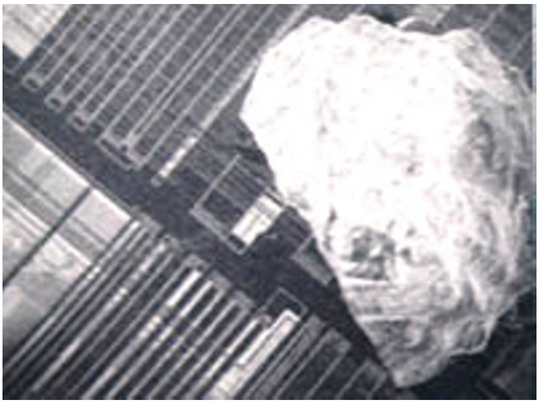


Fig. 3.2. This enlarged image of a grain of salt on a piece of a microprocessor should give you an idea of how small and complex a microprocessor really is. [From http://www.intel.com/education/images/manufacturing/salt.jpg. Reprinted with permission from Intel Corporation.]

Cleanrooms having dust is only half the problem, the other major cause of contamination are the workers themselves. The workers in the cleanroom must wear special uniforms to minimize the introduction of additional contaminants, such as hair, skin flakes, or worse outside world dirt. This protective clothing is made from a non-linting, anti-static fabric and is worn over street clothes. The final step in minimizing particulate based fabrication problems, wafers are chemically cleaned to remove any particles that may have adhered to the surface. This is to promote adhesion of the photoresist to the surface.

3.2.2. Positive and negative photoresists

In photolithography, the resist is called photoresist and it can be of either of the two types: positive or negative photoresist, depending on its chemistry which determines its property when exposed to light. The photoresist is a photosensitive material used to transfer the image from the mask to the wafer surface. The quality of the resist plays an important role in the image transfer to a semiconductor wafer. Resists are generally required to maintain a usable adhesion, uniformity, etch resistance, thermal stability, and long shelf life. Both positive and negative resists are made of complex organic molecules containing carbon, oxygen, nitrogen, and hydrogen. They can be more or less easily dissolved using a developing solution, depending on the amount of light they have been exposed to.

Let us illustrate the difference between positive and negative photoresists by considering the example of a silicon wafer with a thin silicon oxide layer and coated with a layer of photoresist, as shown in Fig. 3.3.

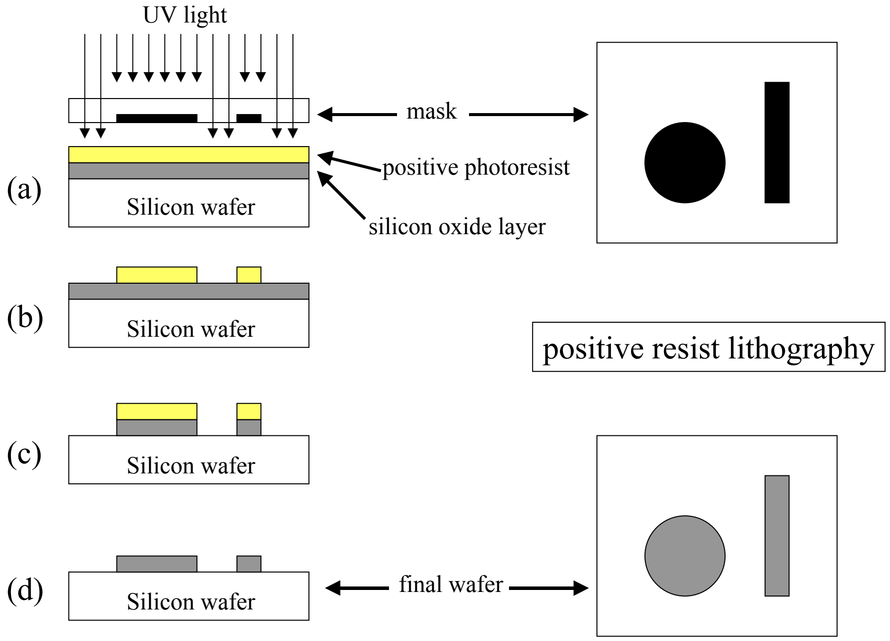


Fig. 3.3. Positive resist photolithography process sequence. When using positive resist, the exposed regions are dissolved in the developing solution, while the unexposed areas remain intact. (a) The positive photoresist is exposed using a source of intense ultraviolet light. (b) The wafer is removed from the alignment station and areas exposed are dissolved in a solution. In the steps illustrated in (c) through (d), the etch-resistant property of the resist is used in the etching of silicon dioxide in the regions which are not protected by the remaining photoresist.

In Fig. 3.3(a), the positive photoresist is exposed using a source of intense ultraviolet light such as a mercury arc lamp which alters its chemical bonding to make it more soluble where it has been exposed. The wafer is removed from the alignment station and developed (Fig. 3.3(b)). The exposed regions of the positive photoresist layer are dissolved in the developing solution, leaving the unexposed areas intact. In other words, for a positive photoresist, the light from the exposure step increases the solubility of the resist in the developing solution by depolymerizing the resist material. Following this image-transfer step to the photoresist, the image needs to be transferred to the underlying layers. In the steps illustrated in Fig. 3.3(c) through (d), the etch-resistant property of the resist is used in the etching of silicon dioxide in regions which are not protected by the remaining photoresist. If properly selected, the etchant will remove the layer of silicon dioxide but will not etch the underlying silicon or the layer of photoresist, as shown in the figure. The result of the photolithographic process is shown in Fig. 3.3(d) where after the layer of resist has been removed, only the patterned layer of silicon dioxide is left.

By contrast, when using a negative photoresist, it is the unexposed regions which are dissolved in the developing solution, leaving the exposed areas intact. In other words, in this case, the light from the exposure step causes polymerization to occur in the resist, reducing its solubility in the developing solution. This is illustrated in Fig. 3.4(a) and (b). The remaining sequence of steps is similar to the previous one and is illustrated in Fig. 3.4(c) through (d).

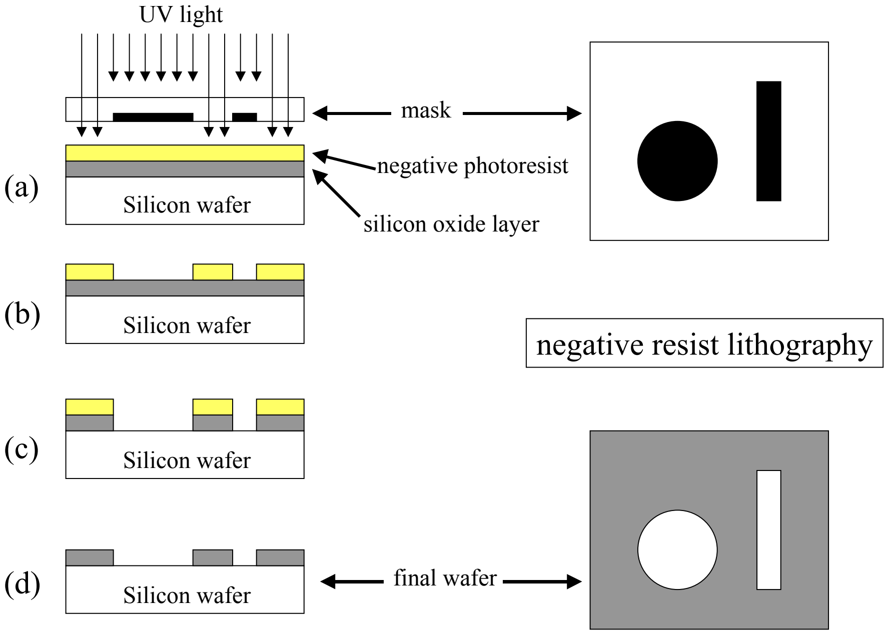


Fig. 3.4. Negative resist photolithography process sequence. When using negative photoresist, the unexposed regions are dissolved in the developing solution, while the exposed areas remain intact. The sequence of steps is similar to that of Fig. 3.3.

The choice of a positive or negative photoresist is determined by the subsequent sequence of processing steps which need to be performed.

For either type of photoresist, the following requirements must be met: it must adhere well to the wafer surface, its thickness must be uniform across the wafer and must be predictable from wafer to wafer, it must be sensitive to light so that it can be patterned, and it must not be attacked by the etchant for removing the substrate material. The photoresist film is first applied to the wafer surface in a yellow light cleanroom environment. In spinning the photoresist, a small puddle of resist is first dispensed onto the center of the substrate, which is attached to a spindle using a vacuum chuck as shown in Fig. 3.5.

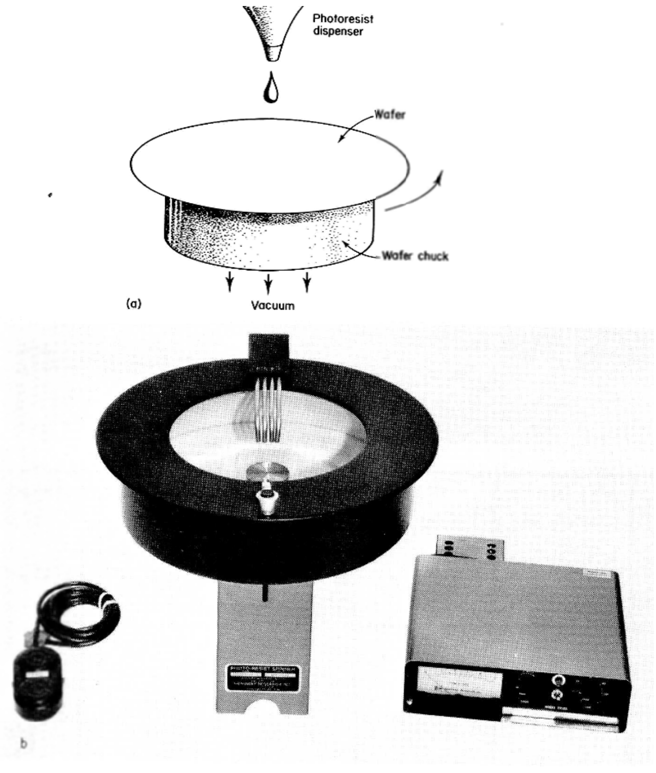


Fig. 3.5. An example of photoresist spinner system. The spinner and its control equipment are shown in (b). The top diagram (a) illustrates the spinning process where a wafer is firmly maintained onto a wafer chuck by pulling a vacuum between them. The wafer/chuck block is then spun and a drop of photoresist is dispensed at the center of the wafer from where it coats the wafer with a thin resist film. [Reprinted with permission from Headway Research, Inc.]

The spindle is then spun rapidly, rotating the substrate at several thousand revolutions per minute for a certain period of time. The formula commonly relating spin speed to final thickness is:

$$\text{Eq. (3.1)} \qquad z = \frac{\kappa_r P}{\sqrt{\omega}}$$

where z is the final resist thickness, P the percentage of solids in the resist, ω the angular velocity of the spinner, and κ_r is a constant that is different for each resist.

For example, for a given spinner and photoresist formulation, we know that a rotation speed of 4000 rpm gives a resist thickness of 0.7 µm. In order to increase the thickness to 0.8 µm, the proper rotation speed ω must satisfy the relation: $\sqrt{\frac{\omega}{4000rpm}} = \frac{0.7}{0.8}$, so that $\omega = 3062$ rpm.

Once the photoresist is applied a prebake, or a softbake, step is performed. Softbaking is used to remove the solvents present in the resist and to improve adhesion. Solvents left in the resist film from poor softbaking will cause a degraded image to be transferred to the wafer surface, because such solvents will be attacked by the developer and cause portions of the resist that are to remain to be removed. Properly baked wafers will have resist that has the proper amount of resins and photosensitizers (positive resist) or inhibitors (negative resist) as determined by the manufacturer. Once the prebake is completed the photoresist-covered wafer is then ready for mask alignment and exposure.

3.2.3. Mask alignment and fabrication

The word photolithography may be loosely defined as "printing with light", which is an accurate description of the heart of this processing step. The manufacture of semiconductor devices and integrated circuits consists of multiple passes through photolithography steps. Each time, it defines the region where the subsequent processing step, e.g. doping introduction, oxidation, metallization, will have its effect. These multiple passes must be aligned using simple marks to help either a computerized aligner or fabricator align the new mask with previous mask step. If masks are not aligned the whole wafer will be dead because the layers of the multiple processes will not be aligned and contacts will not work.

In photolithography, it is first necessary to produce a mask or transparency of the pattern required. Mask making begins with a large-scale layout or artwork which is then photographed by large camera to reduce it down typically more than a thousand times to the exact size on a master plate. Fig. 3.6 shows the sequential steps necessary to create an integrated mask. The master plate is used in a precision step-and-repeat printer to produce multiple sequential images of the layout on a high-resolution photosensitized emulsion glass plate which is later used as the mask in the photolithography process to transfer the layout pattern onto the wafer surface.

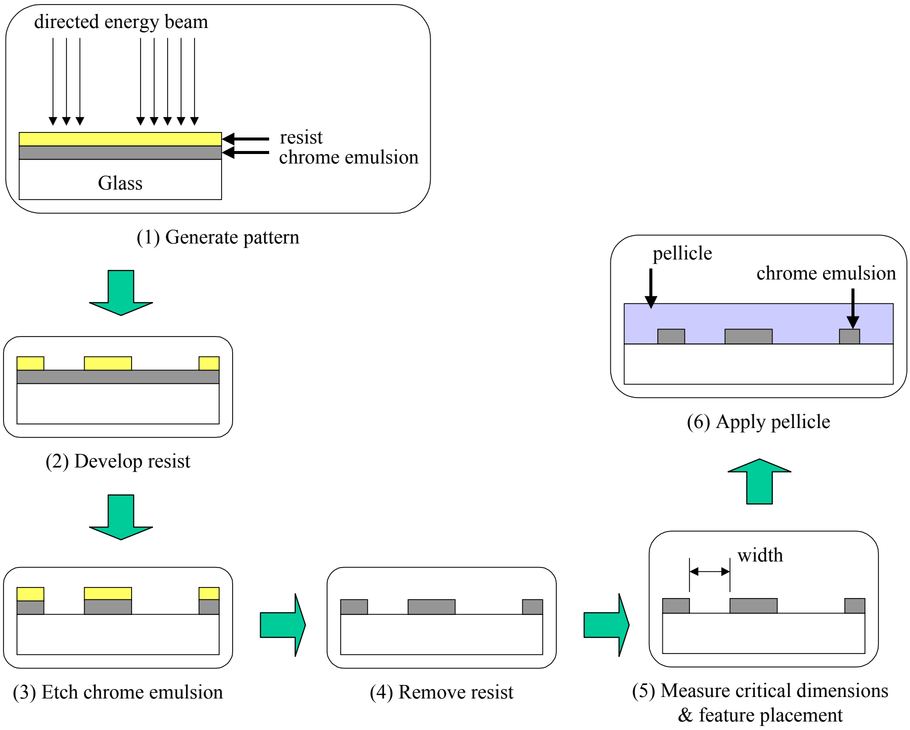

directed energy beam
resist
chrome emulsion
Glass
(1) Generate pattern
(2) Develop resist
(3) Etch chrome emulsion
(4) Remove resist
width
(5) Measure critical dimensions & feature placement
pellicle
chrome emulsion
(6) Apply pellicle

Fig. 3.6. Flow diagram illustrating the realization of a mask which would be later used for photolithography. It begins with a large-scale layout or artwork which is then photographed by large camera to reduce it down typically more than a thousand times to the exact size on a master plate. The master plate is used in a precision step-and-repeat printer to produce multiple sequential images of the layout on a high-resolution glass plate covered with an emulsion and a resist layer. A pellicle barrier layer is provided in order to ensure the integrity of the pattern from particulate.

The emulsion used on the glass plate is susceptible to scratches, wear and tear damage during usage. Alternative materials which withstand wear better than emulsion but are also considerably more expensive, such as chrome and iron oxide, are sometimes substituted for emulsion. Iron oxide masks have the additional advantage of being transparent to the yellow light used to visually align the masks, while being opaque to the intense ultraviolet light used for the exposure of the resist. Visually, each type of mask is a plate of glass with alternate clear and relatively opaque regions as shown in Fig. 3.7.

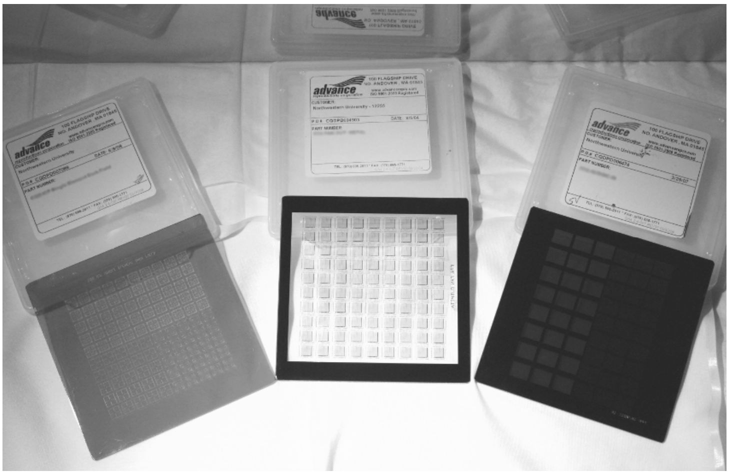


Fig. 3.7. Example of integrated photomasks. Visually, a mask is a plate of glass with alternate clear and relatively opaque regions. From left to right: a semi-opaque iron-oxide dark field mask, a bright field chrome mask, and a dark field chrome mask.

3.2.4. Exposure

As mentioned earlier, the exposure to ultraviolet light determines the property of the photoresist. It is important to know the sensitivity of the resist used and control the amount of light that it receives. In addition it is important to know how the light is passing the mask to hit the surface of the wafer and the photoresist. This affects not only the size of the image made on the wafer but also how long the mask will actually last; an important fact to know due to the length of time and expenses required to make the masks.

There are three main types of printing that are generally used in lithography: contact, proximity, and projection; a schematic diagram of these methods is shown in Fig. 3.8.

In contact printing (Fig. 3.8(a)), the mask is placed in direct contact with the photoresist. This limits diffraction and is the simplest of the techniques to use, but the mask gets worn, the photoresist can be more easily contaminated by residue of the mask, and the mask limits the size of the images produced. Contact printing has been largely replaced by proximity printing (Fig. 3.8(b)); here the mask is held slightly above the wafer. This increases the lifetime of the mask and reduces the potential for contamination but the air gap increases diffraction as well, the image is still determined by the mask dimension size. The third option for optical lithography, projection printing (Fig. 3.8(c)) is now the standard used in fabrication. This is because in this way one can make images smaller than that which appears on the mask. It is made possible because of the use of focus lenses that can shrink the image. This system also allows for larger wafers to have multiple projected images placed upon the wafer by scanning over the entire wafer. This is done using a method known as step-scan processing. While projecting printing allows for reduced contamination like proximity printing, it also able to make smaller images than the mask, and is only limited by the light's diffraction.

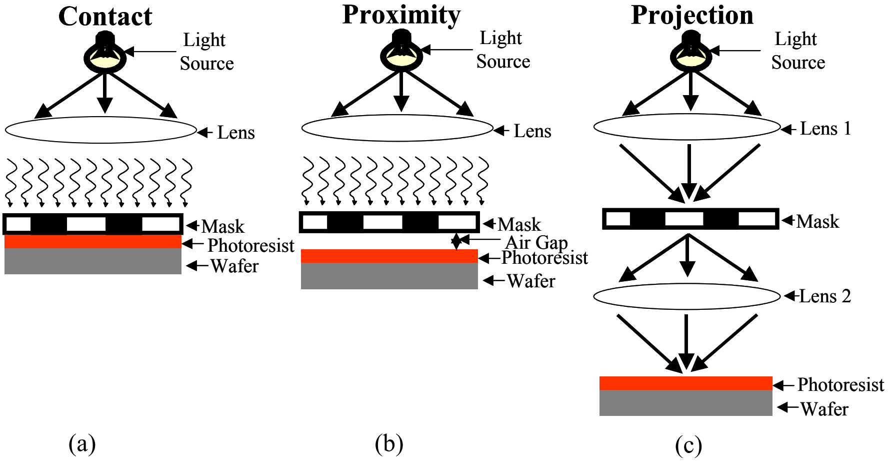


Fig. 3.8. Conceptual drawing of the three main lithographic printing techniques: contact, proximity, and projection.

3.2.5. Development

Indeed, if a photoresist film is underexposed, there is a tendency for the pattern formation to be incomplete and, in the extreme case, to cause a total loss of pattern. For a positive resist which is underexposed, the resist film remains intact after development as shown in Fig. 3.9, whereas for a negative resist it is completely removed. If the film is overexposed, and a positive photoresist is used, then the window openings in the resist are slightly larger than the mask dimension as shown below in Fig. 3.9.

The effects of overexposure also strongly depend on the nature of the photoresist used. For a positive resist, the openings are slightly larger than the mask dimensions as shown below in Fig. 3.10(a). This is due to scattered light penetrating under the mask edges and exposing a small region of film not directly irradiated by the light source. Subsequent etching of the underlying silicon oxide film accentuates this enlargement, as is shown in Fig. 3.10(a). With a negative resist, the window tends to be smaller than the mask dimensions (Fig. 3.10(b)). This can be partially compensated later by the undercut during the etching of the oxide film.

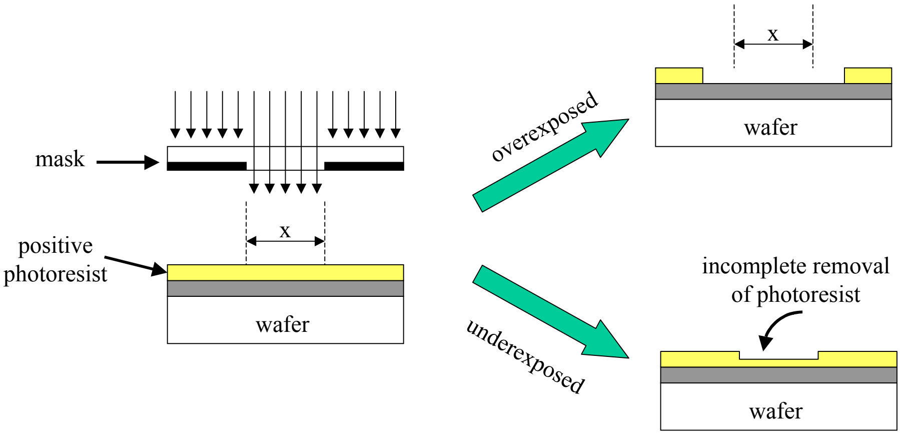

mask
x
positive photoresist
wafer
overexposed
underexposed
x
wafer
incomplete removal of photoresist
wafer

Fig. 3.9. Effect of overexposure and underexposure on the profile of a positive photoresist layer after development.

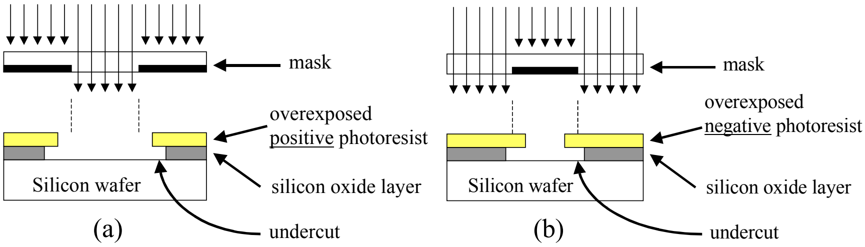

mask
overexposed
positive photoresist
silicon oxide layer
Silicon wafer
undercut
(a)
mask
overexposed
negative photoresist
silicon oxide layer
Silicon wafer
undercut
(b)

Fig. 3.10. (a) Effect of overexposure on the dimensions of a positive photoresist layer: the opening in the resist layer is larger than the size specified by the mask. When using (b) a negative resist, the opening in the resist is smaller than the size specified by the mask. In this latter case, this effect can be balanced by the undercutting which occurs when conducting the wet etching of the silicon oxide layer.

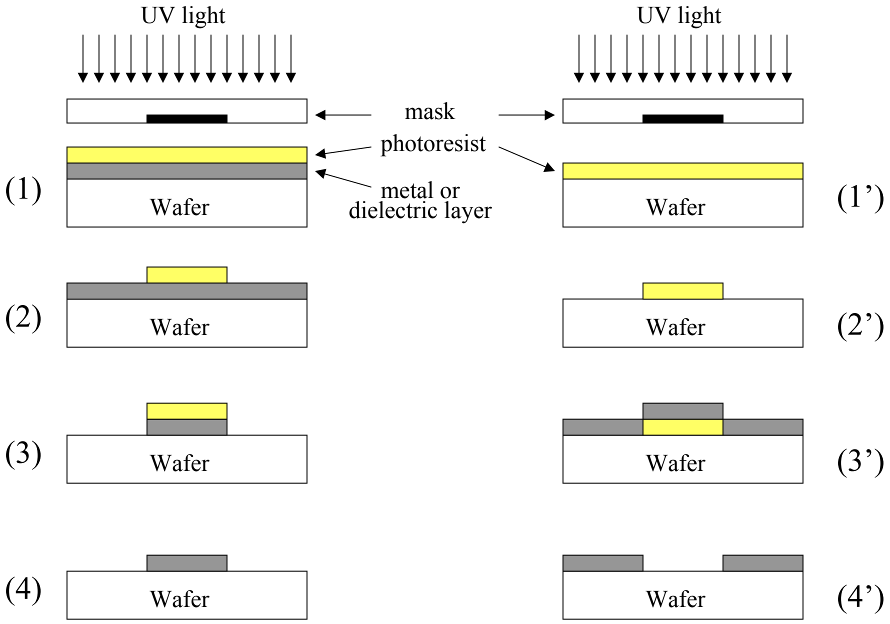
3.2.6. Direct patterning and lift-off techniques

So far, we have described the process of transferring the pattern from the photoresist film to the underlying (oxide) layer through etching, i.e. the areas which were uncovered were etched away. This method is called direct patterning.

In addition to this traditional method, one can use the lift-off technique for depositing and forming patterned metal or dielectric films onto a wafer surface. In this method, the photoresist is patterned first, before an additional (metal or dielectric) film is applied. Subsequently, by removing or "lifting-off" the photoresist, a pattern is achieved in the later deposited film.

To use lift-off process, the order of photolithography is changed in comparison with direct patterning as shown in Fig. 3.11 and the sequence of steps is summarized as follows:

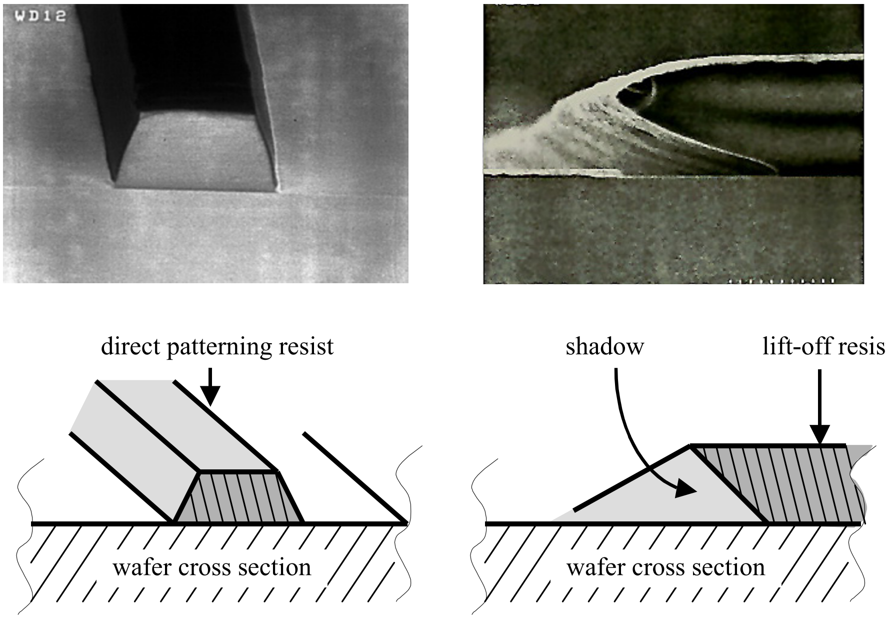
direct patterning sequence	*lift-off sequence*
- Deposit photoresist	- Deposit photoresist
- Softbake	- Softbake
- Alignment/exposure	- Alignment/exposure
- Develop	- Postbake
- Hardbake	- Flood exposure
- Develop	



(a) (b)

Fig. 3.11. Illustration of the (a) direct pattern vs. (b) lift-off techniques using a positive photoresist. In the direct pattern method, the photoresist is applied after the metal or dielectric layer and is exposed (1), developed (2), and the pattern in the photoresist is transferred to the underlying layer through etching (3), before removing the photoresist (4). In the lift-off technique, the photoresist layer is applied, exposed (1') and developed (2') before the metal or dielectric layer is deposited (3'). The photoresist is then lifted off the wafer, taking away with it parts of the metal or dielectric layer (4').

Fig. 3.12(a) and (b) represent cross-section images showing the photoresist profile used in direct pattern and lift-off technique, respectively, following development. The shape of the resist in Fig. 3.12(b) makes it easier to lift-off the subsequently deposited metal or dielectric layer, without peeling it off completely from the surface.



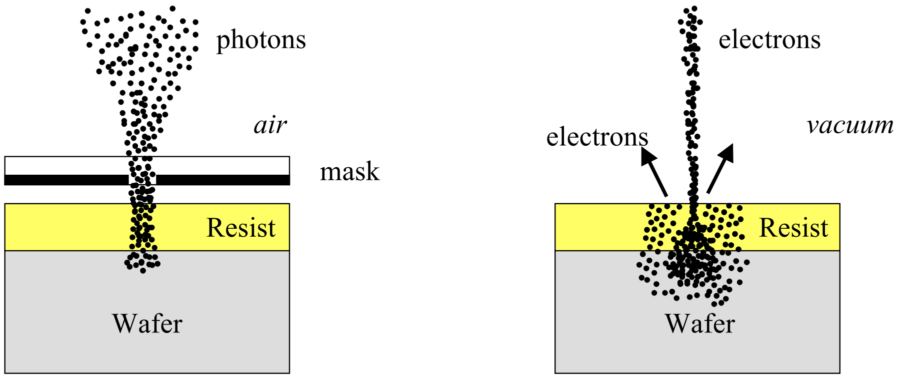
(a) (b)

Fig. 3.12. Cross-section views of a (a) direct patterning resist profile following development, and (b) a lift-off resist profile following development. The bottom diagrams show a schematic diagram illustrating the dissimilar cross-sectional shapes of the direct patterning and the lift-off resists.

3.2.7. Alternative lithographic techniques

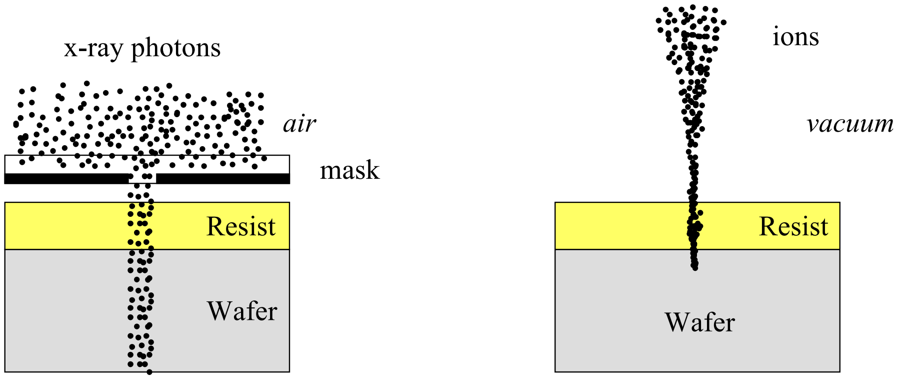
As microelectronic devices shrink in size, alternative approaches have been investigated due to the fact that optical lithography is fundamentally limited by the phenomenon of diffraction. Work is ongoing to develop alternative lithographic techniques that can support Moore's law past this diffraction limit. These next generation lithographies (NGLs) include electron-beam lithography where an electron beam is used as the radiation source. This technique is currently growing in importance. It is currently used for the fabrication of the mask, as well as to define nanoscale features, as small as 30 nm. This technique will be the object of a section of the next section.

Other existing NGLs methods include x-ray lithography which is capable of achieving a few tens of nanometer size features thanks to the use of radiation with a wavelength on the order of 4 up to 50 Å. However, this method requires a complex absorber mask or a thin film support structure. Ion-beam lithography offers patterned doping capability and a very high resolution (~10 nm). All these techniques are schematically illustrated in Fig. 3.13.



Photolithography

Electron-beam lithography



X-ray lithography

Ion-beam lithography

Fig. 3.13. Illustration of different types of lithography. From left to right, and top to bottom: photo-, electron-beam, x-ray and ion-beam lithography. The difference between these techniques resides in the nature of the source used to activate the resist.

While these techniques may seem different from one another, they are fundamentally similar in the fact that they use a photoresist that is activated by a specific energy source. So just like photolithography (Fig. 3.13), they use similar steps as described in this section. The major differences in these types of lithography arise in how they actually interact with the resist. In both photo- and x-ray lithography, the source energy goes through a mask, and the image is projected onto the wafer; this type of lithography is known as an indirect-write pattern. As for ion-beam, or electron beam lithography, the source energy is focused into a fine tip so that the image is directly written onto the surface of the wafer; this known as a direct write pattern.

Two extreme examples of this form of lithography are known as nano-imprint lithography in which a nano-sized stamp is used to impress an image onto the surface or proximal probe lithography (also called Dip-Pen Lithography) where an atomic force microscope tip is used to move material at the atomic scale on the wafer surface. Schematic images of both these techniques are shown below (Fig. 3.14 and Fig. 3.15).

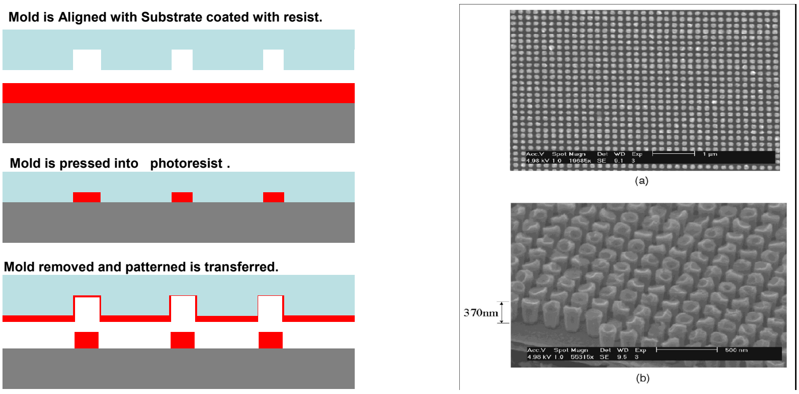

Mold is Aligned with Substrate coated with resist.
Mold is pressed into photoresist .
Mold removed and patterned is transferred.
Acc.V Spot Magn Det WD Exp 1 µm
4.98 kV 1.0 19685x SE 9.1 3
(a)
370nm
Acc.V Spot Magn Det WD Exp 500 nm
4.98 kV 1.0 55315x SE 9.5 3
(b)

Fig. 3.14. In nano-imprint lithography, a mold imprints its image directly onto the wafer surface that is covered by the resist. The resist is developed and the image is then fully transposed. Here, an imprint mask used to make nanopillars is shown with the SEM images of the pillars they produce. [SEM images on the right are reprinted with permission from Journal of Vacuum Science and Technology B Vol. 16, Wu, W., Cui, B., Sun, X.Y., Zhang, W., Zhuang, L., Kong, L., and Chou, S.Y., "Large area high density quantized magnetic disks fabricated using nanoimprint lithography," p. 3826. Copyright 1998, American Institute of Physics.]

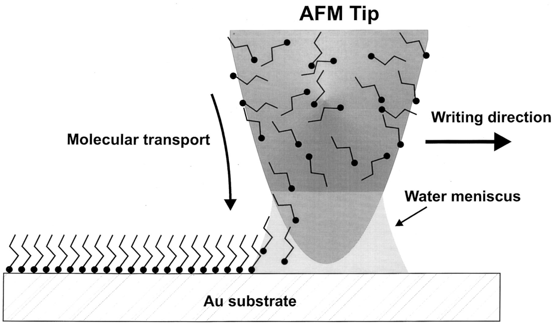

AFM Tip
Writing direction
Molecular transport
Water meniscus
Au substrate

Fig. 3.15. Schematic diagram of dip-pen lithography: either a photoresist is placed down in controlled manner by pacing the AFM tip where it is desired or a molecule is directly placed. [Reprinted with permission from Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A., "Dip-pen nanolithography," Science Vol. 283, p. 661, Fig. 1. Copyright 1999, AAAS.]

3.3. Electron-beam lithography

The resolution of the photolithography process described in the previous section is limited by the diffraction limit of the ultraviolet light source used during the photoresist exposure. As microelectronic devices shrink in sizes to reach the nanometer scale, novel techniques such as electron-beam lithography have been developed and will be the object of this section.

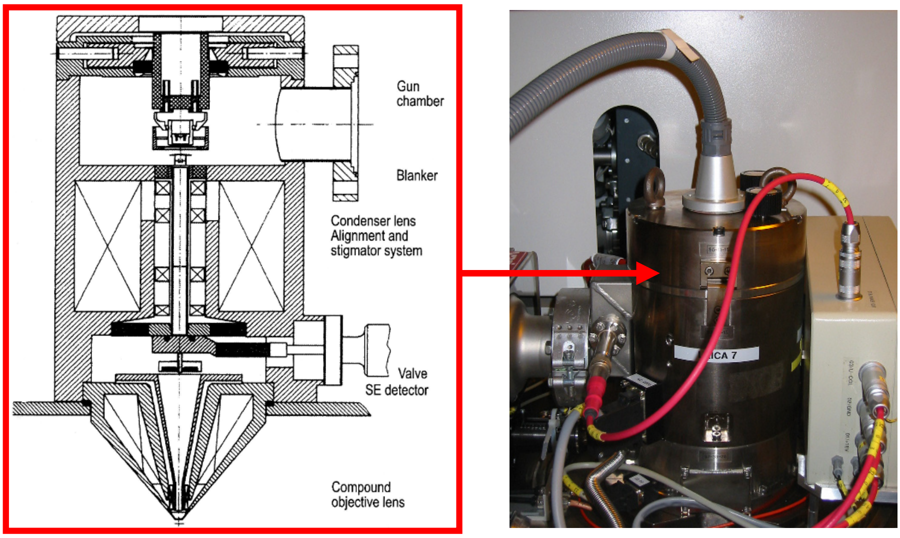
Electron-beam lithography, or EBL, is a special technique for creating extremely fine features of approximately 20 nm in linewidth. This technique was developed in the 1960s and was inspired by the same technology used in scanning electron microscopy where electrons, rather than light, are used to generate an image. In EBL, the electrons strike a thin layer of resist which has been previously applied on the semiconductor wafer. The properties of the resist material are changed by the presence of the electrons that it encounters. Therefore, instead of using a mask to prevent light from passing, a fine electron beam is scanned across the sample in order to form the desired pattern on the resist.

3.3.1. Electron-beam lithography system

The major component in an electron-beam lithography system is the column, a cross-section of which is illustrated in Fig. 3.16. It contains an electron source, two or more lenses, a blanker that can switch the beam entrance on and off, a beam deflector, a stigmator for correcting any astigmatism in the beam, an aperture for helping to define the beam, alignment systems for centering the beam in the column, and an electron detector for assisting with focusing and locating marks on the sample.

The electron source consists of a heated filament which generates a beam of electrons through thermionic emission. The size of the virtual source, its brightness, and energy spread are the three most important characteristics of the electron beam. The source size dictates the amount of demagnification that the beam must undergo in order to affect the target in a small area. The brightness, measured in amperes per square centimeter through the column, must be high enough to sufficiently affect the resist. The energy spread determines the tendency of the electrons to move outward from the direction of the main beam. There is always some energy spread due to the electric field interaction between the electrons in the beam but this effect can be corrected using apertures. A change in the electron energy will cause the virtual emitter source to change its position slightly too. Provided vacuum is maintained at the specified levels, the electron source may have a lifetime of more than 2000 hours.

The lenses in the column help aim the beam toward the appropriate area of the target. They do not, however, scan the beam across the sample in order to etch patterns on the resist; as that task is left to the deflectors. There are two types of lenses that are commonly used. Electrostatic lenses employ electric fields in order to have an effect on the beam. However, they suffer from both spherical and chromatic aberrations. Spherical aberrations occur when the outside edges of the lens cause the beam to focus more strongly than the inside areas do. Chromatic aberrations are observed when the lens affects differently electrons that have different energies. Both of these effects can be reduced by drastically reducing the size of the beam so that it only passes through the center area of the lens, but this technique reduces the beam current. Magnetic lenses are more commonly used because they cause less aberration.



(a) (b)

Fig. 3.16. (a) Schematic diagram of a low-voltage column, containing the condenser and objective lens systems, as well as the beam blanker, the alignment and deflector systems, and an electron detector. (b) Photograph of the low-voltage column inside an actual electron beam lithography system. [Reprinted with permission from Leica Microsystems Lithography GmbH.]

Apertures are small holes through which the beam passes as it travels down the column. There are three types of apertures, depending on their diameters. Let us assume that the diameter of the main portion of the beam (excluding stray electrons) is D_B. A spray aperture, which stops stray electrons but does not affect most of the beam, would have a diameter $D_A > D_B$. A blanking aperture, which is used to turn the beam on and off, has a diameter of $D_A = 0$. A beam limiting aperture, which is reduces the beam diameter in order to improve the resolution, would have a diameter $D_A < D_B$.

The method of appropriately scanning the electron beam in a precise pattern is known as deflection. Similar to lenses, deflectors generate electric or magnetic fields that are used to affect the direction of the electron beam. Electrostatic lenses are more commonly used than magnetic lenses for this purpose because they react faster to the system's demands to scan the beam across the target surface. This difference in speed arises from the existence of inductive magnetic coils necessary to create a magnetic field. Difficulties are encountered when the electron beam interacts with its solid target. Although the beam may be extremely small when it first hits the resist material, the interaction of the beam's electrons with the material causes what is known as forward scattering. For example, an electron that penetrates 1 μm deep into the resist layer when impinging at a 90° angle would travel not only 1 μm down, but also 1 μm laterally when deflected at a 45° angle. This phenomenon results in an error in the actual area of the material which is hit by electron. Using the thinnest possible resist layers can reduce forward scattering. These errors can also be minimized with a technique called dose modulation in which small, isolated areas receive a less intense electron beam current dose than large areas.

Underneath the column, a chamber contains a stage used to load and unload the sample. In addition, a vacuum system is used to maintain an appropriate vacuum level throughout the machine and during the load and unload cycles. A computer controls the EBL system and handles such diverse functions as setting up an exposure job, loading and unloading the sample, aligning and focusing the electron beam, and sending pattern data to the pattern generator.

Electron-beam lithography is becoming an increasingly common alternative to photolithography because of its near-atomic resolution (~20 nm) capability, its flexibility in that it works well for a variety of semiconductor materials and an almost unlimited number of patterns. However, electron-beam lithography is more than ten times slower than photolithography. EBL systems are also expensive and are complex pieces of equipment which require frequent maintenance and adjustment. These limitations keep EBL from becoming the semiconductor industry's lithography standard.

3.3.2. Electron-beam lithography process

Electron-beam lithography uses resists known as polymethyl methacrylate or PMMA which are some of the highest resolution resists available. The PMMA is purchased in two high molecular weights forms (496 K or 950 K) in a casting solvent such as chlorobenzene or anisole.

Fig. 3.17 briefly illustrates the electron-beam lithography process, which is similar to that of conventional photolithography, including the steps such as resist spinning, exposure, development, and pattern transfer.

To avoid charging effects during electron-beam exposure, the wafer is often first coated with a thin (3~30 nm) indium-tin-oxide (ITO) or chrome layer. The PMMA resist is spun onto the wafer and baked at 170~200 °C for 30 minutes. The electron-beam exposure breaks the resist polymer into fragments that can then be dissolved by a developer solution. There are several types of developer solutions with different strengths, such as MIBK which is a 1:2 solution of (4-methyl-2-pentanone):(2-propanol) or IPA which is simply 2-propanol. MIBK alone is a strong developer and dissolves some of the unexposed resist too, while IPA is a weaker developer. Therefore, mixing them with the appropriate proportions results in a higher contrast or a higher sensitivity for the resist. These two parameters will be defined shortly below. For example, a mixture of 1 part MIBK to 2 parts IPA produces very high contrast but low sensitivity whereas, for a mixture of 1 part MIBK to 1 part IPA, the sensitivity is improved significantly with a small loss of contrast.

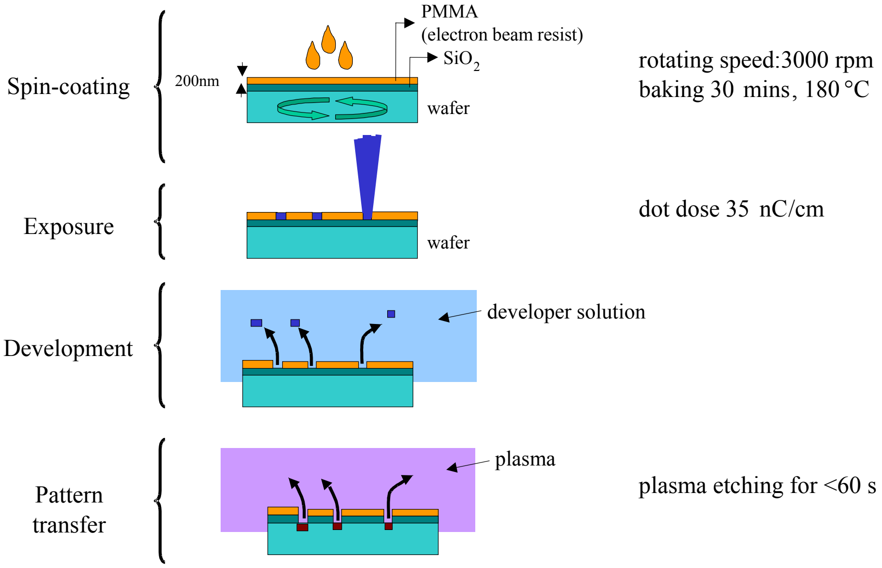

Spin-coating
200nm
PMMA
(electron beam resist)
SiO_2
wafer
rotating speed:3000 rpm
baking 30 mins, 180 °C
Exposure
wafer
dot dose 35 nC/cm
Development
developer solution
Pattern transfer
plasma
plasma etching for <60 s

Fig. 3.17. Outline of a typical electron-beam lithography procedure: a PMMA resist film is spun onto a wafer, the resist is exposed to the electron beam following the specified design, the exposed resist is then developed and areas which have been exposed are dissolved, The pattern is transferred on the wafer through plasma etching. The resulting features can be as small as a few nanometers in width.

3.3.3. Parameters of electron-beam lithography

Several parameters need to be understood and determined for electron-beam lithography. For a given set of resist conditions, including the nature of the resist, its thickness, the nominal dose D_{nom} corresponding to a given electron energy level can be defined as the minimum dose required to ensure full dissolution (when using positive resist) or total non-solubility (when working with negative resist) of the resist in all places that were exposed to that electron beam.

If we chose to expose a uniform positive resist to a range of doses, develop the pattern and then plot the remaining average resist thickness (expressed in terms of film retention as a percentage) versus dose, we would obtain the graph shown in Fig. 3.18(a).

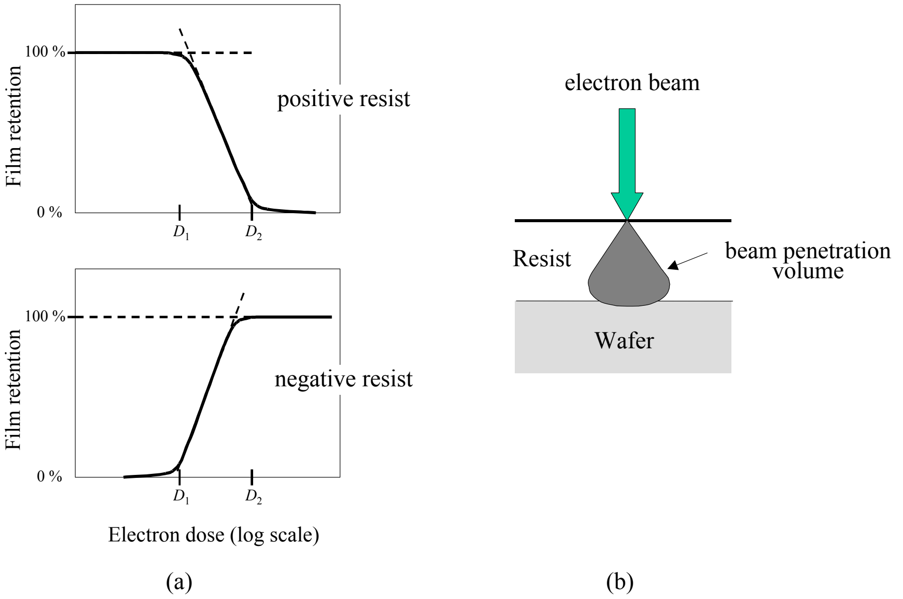


Fig. 3.18. (a) Graphical plots of the resist film thickness as a function of the electron exposure dose for a positive and for a negative resist. For a positive resist, D_1 is the largest dose at which the film remains intact while D_2 is the smallest dose at which the entire film is dissolved. For a negative resist, the meaning of these parameters is reversed. These quantities are used to determine the sensitivity and the contrast of the resist. (b) Illustration of the penetration of the electron beam into the resist film.

In the case of a positive resist, D_1 is the largest dose at which the film remains intact while D_2 is the smallest dose at which the entire film is dissolved. The sensitivity of the resist is defined as the point at which the entire resist is removed. We can also define the contrast γ of the resist as:

Eq. (3.2) $$\gamma \equiv \left| \log\left(\frac{D_2}{D_1} \right) \right|^{-1}$$

The same expression is valid for the contrast of a negative resist, but the meaning of D_1 and D_2 are swapped as shown Fig. 3.18(a). To obtain such graphs, it is essential that the electron beam penetrates the resist completely, as shown in Fig. 3.18(b). To ensure this, the resist must be thinner than the penetration depth of the electrons into the resist material, which is determined by the electron beam energy.

To determine the nominal dose for a given set of resist conditions, especially a given resist thickness, it is recommended to use the "Taxi checker" pattern as illustrated in Fig. 3.19. The inner part of the pattern contains two lines having twenty square sections with an edge length of 5.0 µm. Each square receives a set dose. The applied dose is raised incrementally toward the right, as shown in Fig. 3.19. The dose variation is achieved by varying the exposures time or dwell time, and begins with an initial dose D, hopefully close to the desired nominal dose. This selected dose corresponds to a time t_{dwell} calculated according to:

Eq. (3.3) $$D = \frac{I_{probe} \times t_{dwell}}{(SSZ)^2}$$

where D is the applied dose expressed in C·cm^{-2}, I_{probe} is the probe current at the target level (pA), t_{dwell} is the exposure time per image spot (s) and SSZ is the step size in nm.

The first square in the upper lines is exposed for one tenth of the nominal time, which corresponds to ten percent of the nominal dose. Proceeding toward the right, every further square is given a 10% higher dose. This is accomplished by incrementing the exposure time by 10% of t_{dwell}. Exposure of the lower line begins with the first square being subjected to a dose of 110% of the initially selected dose. The two checker lines thus cover a range between 10 and 200% of the originally selected dose.

The following sequence of steps is conducted to determine the nominal dose: (i) coat the wafer with PMMA resist with the desired appropriate thickness; (ii) prepare for exposure by optimizing the probe current, calibrating the main deflector unit and the beam tracking unit; (iii) expose the taxi checker patterns, if there is absolutely no initial information about the sensitivity of the resist system to be investigated, then several taxi checkers patterns should be exposed; and this should be done with doses selected such that the largest possible dose area is covered; repeat the same exposure procedure on several other wafers if necessary; (iv) develop the resist by selecting a development time slightly shorter than would be needed for complete development; (v) inspect the resist image under an optical microscope and determine the dose value; (vi) develop the same resist again if possible, for example for 20 additional seconds; (vii) inspect resist image under an optical microscope and determine appropriate nominal dose; (viii) repeat the development and dose determination until the completely developed square no longer shifts toward higher doses, i.e. toward the right, and an additional development and determination step only yield a general change in contrast.

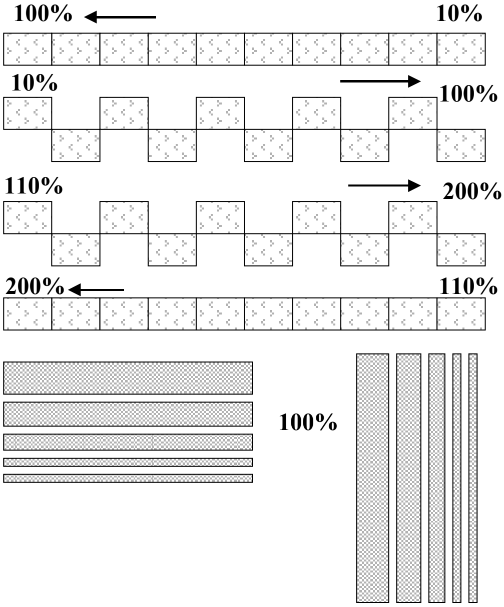


Fig. 3.19. Taxi checker pattern, commonly used to determine the nominal dose for a given set of resist conditions. The inner part of the pattern contains two lines having twenty square sections with an edge length of 5.0 μm. Each square receives a set dose. The applied dose is raised incrementally toward the right.

3.3.4. Multilayer resist systems

In electron-beam lithography, it is often necessary to employ simultaneously two or more different types of resists to achieve a specific lithographic objective when an enhanced undercut is needed for lifting off a metal layer, when a rough surface structure requires planarization, and when a thin imaging top layer is needed for high resolution. A few examples will be given in this sub-section to illustrate this process.

The first example consists of utilizing both a low and a high molecular weight PMMA resist, in a simple bilayer technique as shown in Fig. 3.20.

This technique was patented in 1976 by Moreau and Ting and was later improved by Macki and Beaumont by the use of a weak solvent (xylene) for the top layer of PMMA. The low weight resist is more sensitive than the high weight one, so that it develops more easily and results in an enhanced undercut. This feature is useful when lift-off is required from densely packed layers.

The second example uses PMMA with a copolymer resist and was developed by Hatzakis. This method is often employed when it is necessary to deposit a very thick layer of metal (>1 μm in thickness). A high sensitivity copolymer of methyl methacrylate and methacrylic acid (PMMA-MAA) is spun on top of a PMMA resist layer. The exposed copolymer is soluble in solvents such as alcohol and ethers but insoluble in nonpolar solvents such as chlorobenzene. A developer such as ethoxyethanol/isopropanol is used for the top layer, stopping at the PMMA. Next, a strong solvent such as chlorobenzene or toluene is used for the bottom layer. Through this technique, a larger undercut resist profile is achieved which helps the lift-off for the thick metal layer and has been successfully used in the fabrication of memory arrays.

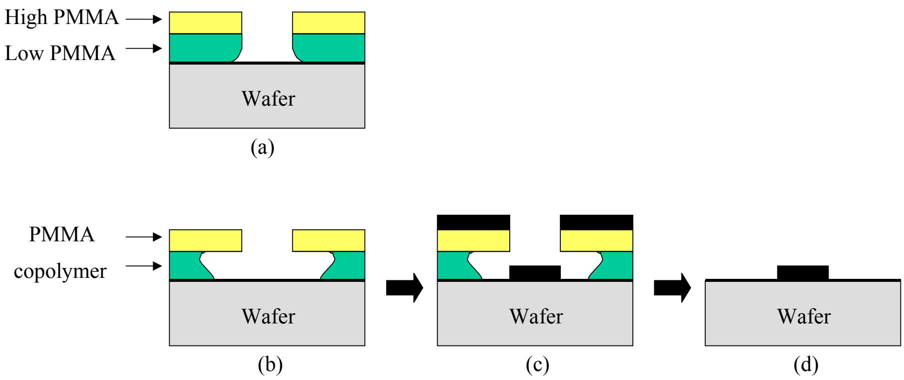


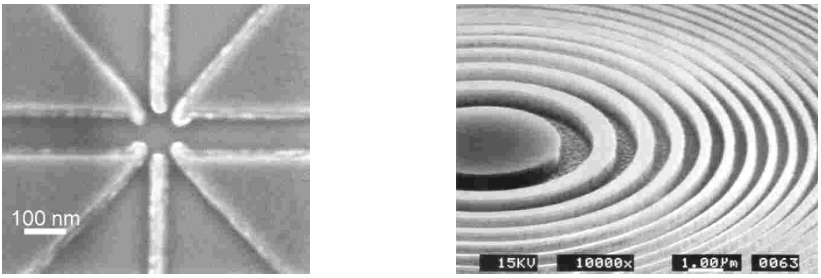
Fig. 3.20. Bilayer electron-beam resist structure: (a) a high molecular weight PMMA is spun on top of a low molecular weight PMMA. The resist is then developed in MIBK:IPA giving a slight undercut. (b) PMMA is spun on top of the copolymer. The resist is developed in MIBK:IPA 1:1 giving a larger undercut. (c) Metal is deposited on top of the resist and (d) is then removed through lift-off.

The third example consists of a trilayer system in which an interlayer is inserted between the two films of the previous bilayer system. Almost any two polymers can be combined in a multilayer system if they are separated by a barrier such as Ti, SiO_2, Al, or Ge. Once the top layer is exposed and developed, the pattern is transferred to the interlayer through dry etching methods (section 3.4) to achieve highly vertical etch profiles. This interlayer would then serve as an excellent mask for the subsequent fabrication of density packed, high aspect ratio resist profiles.

3.3.5. Examples of structures

Electron-beam lithography is by far the most widespread lithography tool for the realization of nanoscale structures. Fig. 3.21 illustrates the resolution and the uniqueness of structures that can be done using this technology. The patterns have linewidths smaller than 100 nm.

The uniqueness of most EBL systems resides in their ability to produce different types of patterns ranging from circular to linear gratings on any type of semiconductor material. This capability is currently used considerably in university research for the development of distributed feedback lasers operating at various wavelengths ranging from 300 nm up to 10 µm.



(a) (b)

Fig. 3.21. (a) Ti/Al gate structure for a SET device generated by electron-beam lithography and lift-off, and (b) a Bragg-Fresnel lens for x-rays exposed in continuous path control mode and etched into Si.

Fig. 3.22 shows an example of linear grating fabricated on top of a ridge quantum cascade laser emitting at 9.0 µm. The grating consists of a first order Bragg grating with a pitch of $\Lambda = 1.42$ µm exposed by electron-beam lithography and etched 0.5 µm into the surface of the 1 µm-thick top cladding layer by reactive ion etching.

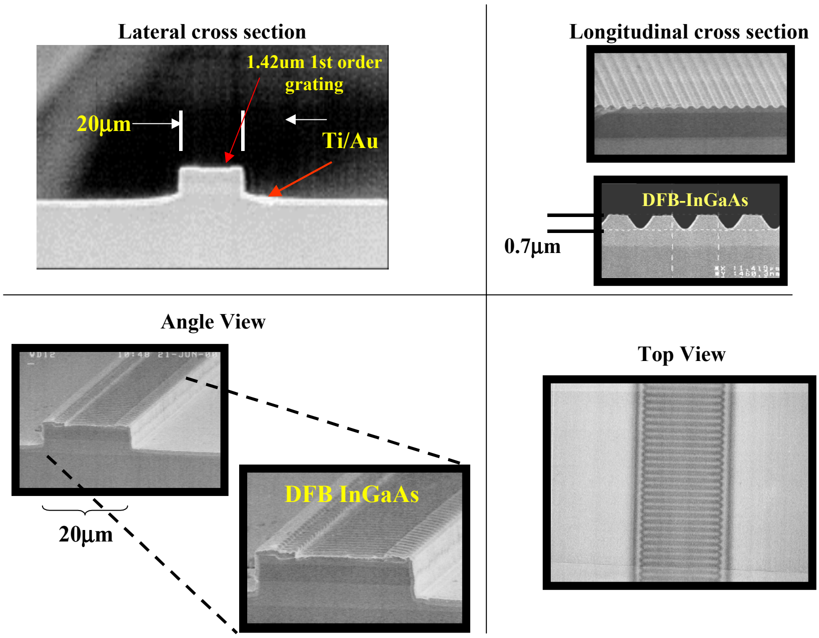


Fig. 3.22. Scanning electron microscopy images of the first order distributed feedback grating for a 9.0 µm quantum cascade laser. The grating is fabricated on the top of the ridge structure using electron-beam lithography.

3.4. Etching

In the previous sections, we have illustrated our discussion with the etching of a layer which had first been covered with a patterned photoresist film, leaving certain selected areas open and others protected. The layer to be etched is generally a dielectric material such as silicon oxide, or a metal used in providing metal contact to the semiconductor. The etching step itself is a complex process which is a function of numerous parameters. For example, the etch can be isotropic, such that the material is etched in equal proportions in all directions, or anisotropic such that one direction is etched more rapidly than any other, or a mixture of both. Several etching techniques can be used and will be described in this section, including wet chemical etching, and dry etching techniques such as plasma etching, reactive ion etching, sputter etching and ion milling.

3.4.1. Wet chemical etching

Wet chemical etching is a mostly isotropic process that etches away in all directions. The process is accomplished by immersing the wafers in an etching solution at a predetermined temperature. An example of solution can be a mixture of hydrofluoric acid and ammonium fluoride. A great variety of wet etch chemistries are available.

One can usually find a solution which is highly selective for the etching of a particular layer while leaving essentially unaffected the adjacent or underlying materials. This material selectivity is an important issue when etching a semiconductor device as it usually contains numerous layers, or when etching a metal contact layer without affecting the underlying semiconductor material.

For most wet etch processes, the material to be etched is not directly soluble in the etching solution, but rather undergoes a reaction with the chemicals present in the solution. The products of this chemical reaction can then be soluble in the solution or can be gaseous. In the later case, the gas can form bubbles which can then prevent the arrival of fresh etching chemical species from reaching the wafer surface to sustain the chemical reaction. This can be a serious problem since the occurrence of these bubbles cannot be predicted. The problem is most pronounced near the edges of the pattern. Mechanical agitation of the wet etching solution can reduce the ability of the bubbles to adhere to the wafer, as well as help sustain the supply of fresh etching reactant.

The advantages of wet etching include its lower cost and the greater versatility of the etching equipment available. Several factors however may affect the quality of wet chemical including: the fact that the photoresist often loses its adhesion to the underlying material when exposed to hot acids, the etching proceeds downward and laterally, thus producing undercutting and broadening lines, and it is difficult to control the etching for submicron geometries. Table 3.1 and Table 3.2 list a few semiconductors, dielectric materials, and metals which are commonly etched in modern microelectronic device fabrication, together with a few wet chemical etching solutions typically used.

Wet etching is also dependent on the crystallographic orientations of the semiconductor crystal, which determines the atomic packing density of the different planes exposed to the etching chemicals. Fig. 3.23 shows the etch planes and profiles when the protective resist is oriented along various directions on a (001) GaAs wafer.

Material	Wet etching solution
Si SiO_2	$HNO_3 + HF + H_2O$ (5:3:3) HF $HF + H_2O$ BHF (H_2O+HF+NH_4F)
GaAs	$H_2SO_4 + H_2O_2 + H_2O$ $NH_4OH + H_2O_2$ (pH=7)
InAs	$H_2SO_4 + H_2O_2 + H_2O$ Br_2 + Methanol $HF + HNO_3 + CH_3COOH$
GaSb	Br_2 + Methanol $HF + HNO_3 + CH_3COOH$ $CH_3COOH + HCl + HNO_3 + Br_2$
InGaP	HCl
Si_3N_4	H_3PO_4 (T = 160~180 °C)

Table 3.1. A few semiconductors and dielectric materials commonly encountered in microelectronic device fabrication and the common wet etching solutions used.

Metal	Wet etching solution
Au (gold)	$KI + I_2 + H_2O$
Pb (lead)	$H_2O_2 + CH_3COOH$ $HNO_3 + H_2O$
Ti (titanium)	$H_2SO_4 + HNO_3 + H_2O$ (polishing etch) HF
AuZn	$KI + I_2 + H_2O$
Ni (nickel)	$HNO_3 + CH_3COOH + H_2SO_4$ HCl $H_2SO_4 + H_2O$ (Electrochemical polish)
Al (aluminum)	$H_3PO_4 + HNO_3 + CH_3COOH + H_2O$

Table 3.2. A few metals and their wet etching solutions.

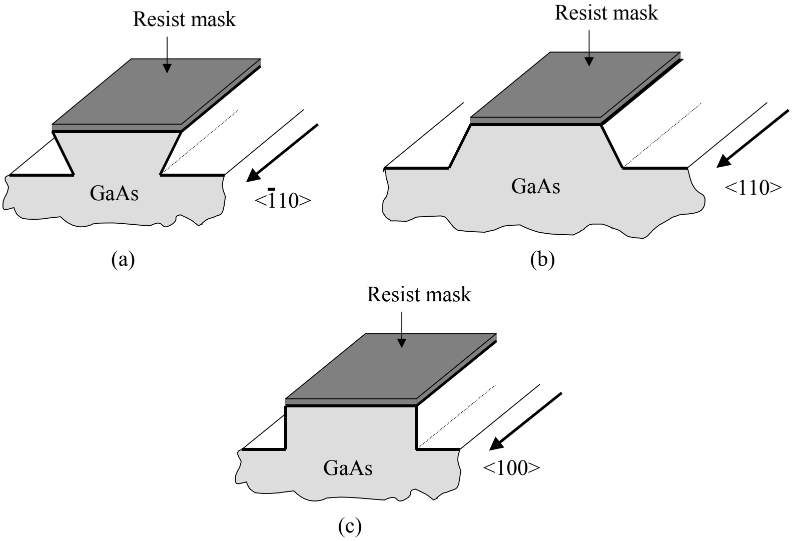


Fig. 3.23. Etch profiles of a (001) oriented GaAs crystal, obtained with most GaAs etchants as a function of the in-plane crystal orientation. When the side of the protective resist mask is oriented in the (a) $<\bar{1}10>$, (b) <110>, and (c) <100> directions.

3.4.2. Plasma etching

By contrast to wet etching, dry etching is not performed in a solution but rather in a gaseous environment. It either consists of plasma driven chemical reactions and/or energetic ion beams aimed at removing the material. Dry etching is commonly used to obtain highly anisotropic etch profiles as the one shown in Fig. 3.24. Some of the advantages of dry etching over wet etching are its greater control at a reduced cost, its substantial directionality i.e. high anisotropy, its effectiveness to reduce the undercutting of masking patterns, and the possibility to precisely etch smaller geometry features. There exist a variety of dry etching techniques including: plasma etching, reactive ion etching, sputtering etching, and ion milling. In this sub-section, we will describe the plasma etching method.

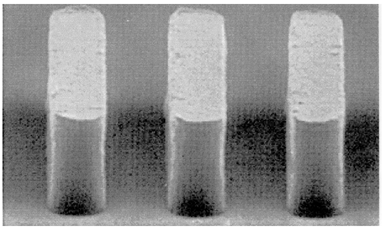


Fig. 3.24. Scanning electron image of a highly anisotropic etch profile obtained using dry etching techniques.

Plasma etching refers to any process in which a plasma, or a gas of charged particles, generates reactive enough species that they serve to chemically etch or physically remove material in the immediate proximity of the plasma. The wafers masked with a photoresist are placed in a vacuum chamber system. A small amount of reactive gas plasma, for example of oxygen, chlorine, or fluorine, is allowed into the chamber. An electromagnetic field is then applied to obtain a directional beam of excited ions and the material that is not protected by the photoresist is etched away by the excited ions. The key to plasma etching is the ability to couple the electromagnetic energy into the reactive species while not heating the rest of the gases in the chamber. Fig. 3.25 and Fig. 3.26 show two popular types of plasma etching reactors: a barrel reactor or a planar reactor.

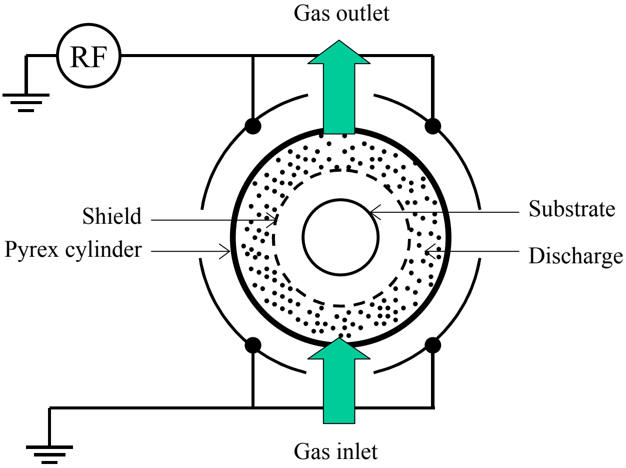


Fig. 3.25. A typical barrel reactor. The plasma is excited using inductive or capacitive electrodes outside of the quartz chamber. The substrates are held in the vertical position by a slice holder and are immersed in the plasma with no electrical bias applied.

In barrel systems, the plasma is excited using inductive or capacitive electrodes outside of the quartz or glass cylindrical chamber. The substrates are held in the vertical position by a slice holder and are immersed in the plasma with no electrical bias applied.

The planar system consists of two flat and parallel electrodes of the same size. The substrates are placed flat on the lower electrode which is also used as a heating stage. Electrons are created in the plasma by the dissociation of atoms into ions. Since they have a greater mobility than the positive ions, they move from the plasma onto the electrode surfaces, thus giving them a negative charge with respect to the plasma. This results in an electric field across the plasma sheath, between the plasma and the electrodes. This field then causes the ions at the edge of the plasma sheath to be accelerated toward the electrodes.

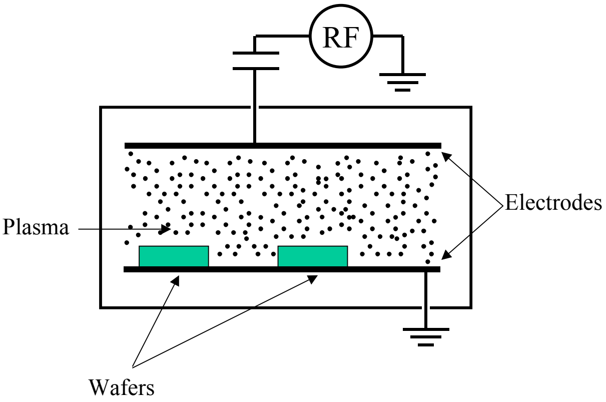


Fig. 3.26. A typical planar reactor, consisting of two flat and parallel electrodes of the same size. The substrates are placed flat on the lower electrode which is also used as a heating stage. The electric field which appears between the plasma sheath and the electrodes causes the ions at the edge of the plasma sheath to be accelerated toward the electrodes, thus impinging on the wafers.

Because of the geometry of the planar reactor, the ions are accelerated perpendicularly to the electrodes except near the outer radius where deviations can lead to corresponding distortions in the etch profile. This perpendicular impingement of energetic ions makes the anisotropic etching possible.

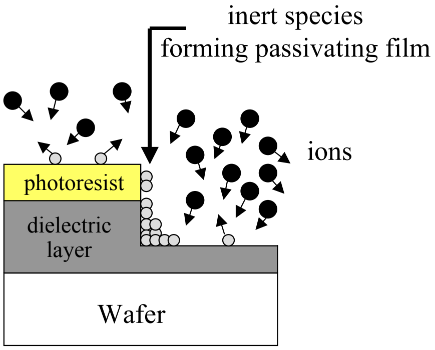


Fig. 3.27. Schematic diagram of an anisotropic plasma etch, showing the formation of a passivating film on the sidewall from the products of the chemical reactions which occur during the etch.

Fig. 3.27 illustrates an anisotropic plasma etch. This process can also lead to the formation of a passivating film on the vertical sidewalls. For example, when a fluoride compound is used in the etching chemistry, e.g. carbon tetrafluoride or CF_4, a fluorocarbon film deposits on all surfaces. Since the ions mostly follow a vertical path, there is little ion bombardment on the sidewalls and the fluorocarbon film can accumulate there, as the etching proceeds. The nature of such a film depends on the plasma conditions.

Table 3.3 lists the advantages and disadvantages that are commonly associated with in plasma etch using either type of reactors. It is important to be aware that the plasma etch rate for a process is measured for a given set of process conditions, from which the duration needed to etch a layer with a particular thickness can subsequently be determined.

	Advantages	Disadvantages
Barrel reactor	- good to remove resist	- poor uniformity due to gas flow RF fields - etch rates tend to increase from the center of wafer to the edges
Planar reactor	- good uniformity - suitable for selective etching through masking patterns	- ionic bombardment can damage wafer surface layers - can lead to significant undercutting

Table 3.3. Advantages and disadvantages in plasma etch using barrel or planar reactors.

Plasma etching provides a poor reproducibility because subtle changes in the etch process or in the film properties can result in poor uniformity or rough surfaces. Fig. 3.28 is an example of a planar reactor plasma etch where the surface of the sample has been damaged due to an increase in the energy of the bombarding ions.

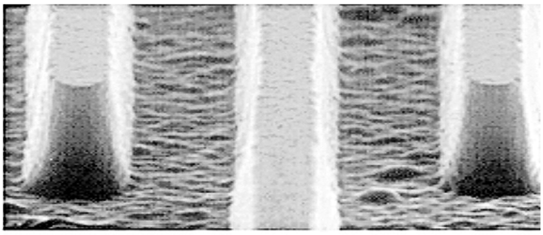


Fig. 3.28. Scanning electron microscope image of a damaged semiconductor surface obtained when excessive ion bombardment is used. The damage can be seen from the roughness of the surface.

3.4.3. Reactive ion etching

Reactive ion etching (RIE) is very similar to plasma etching and a clear distinction is difficult to make. In the RIE process, the emphasis is primarily put on the directionality of the etch. The etch parameters such as the pressure and the configuration of the etching equipment are modified to ensure a directional ion bombardment. A typical RIE system is illustrated in Fig. 3.29.

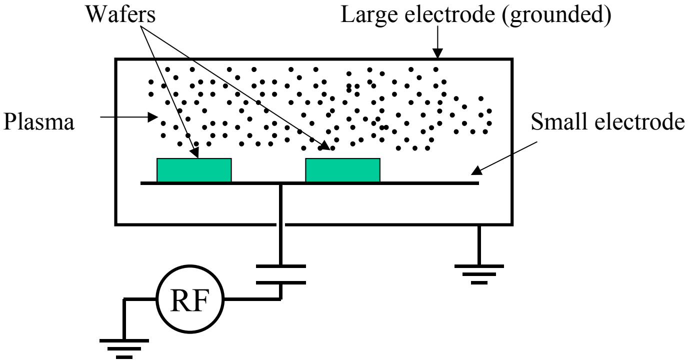


Fig. 3.29. A typical RIE etching system which is similar to plasma etching. It includes a chamber under vacuum, the wafers are placed on a small electrode and the inside walls of the chamber constitute the ground electrode.

By contrast to plasma etching, RIE systems operate at much lower pressures: 0.01 to 0.1 Torr. The advantages of RIE are its highly anisotropicity and directionality, but its disadvantages are that the stage needs to be cooled in order to resist the temperature rise.

3.4.4. Sputter etching

Sputter etching is a purely mechanical process in which energetic ions from the plasma of inert gases, such as argon, strike the wafers and physically blast atoms away from the surface. The sputtering technique and the reactive ion etching are both carried out with the wafer placed directly on an electrically powered electrode in contact with the plasma, as shown in Fig. 3.30. In this configuration, the ions impinge on the sample at near normal incidence. In sputter etching, the chamber is maintained at a low vacuum (10^{-2} Torr) and, as it contains the plasma discharge, it is exposed to ultraviolet radiation, x-rays, and electrons as part of the plasma environment.

The advantages of sputter etching include its high anisotropy, whereas one of its disadvantages is its poor chemical selectivity: all materials are nearly etched at the same rate in this process.

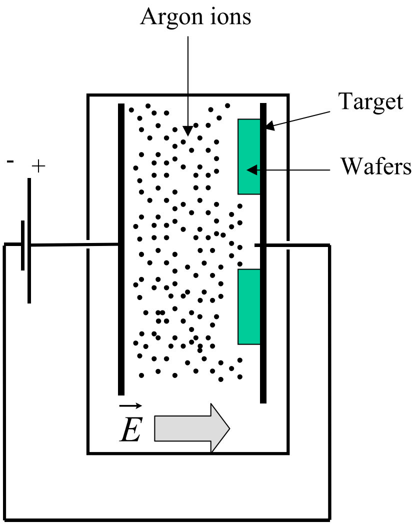


Fig. 3.30. Schematic diagram of a sputter etching system. Energetic ions from the plasma of inert gases are accelerated between two electrodes in a chamber under vacuum and strike the wafers to be etched which are placed directly on one of the electrodes in contact with the plasma.

3.4.5. Ion milling

Ion milling is also a purely mechanical process in which the incoming ions are energetic enough to sputter material from the surface of the wafer. Inert gases are also generally used here as well. One key parameter in this process is the sputtering yield which is defined as the number of sputtered atoms per incident ion. This quantity depends on the material being etched, the nature of the impinging ions, their energy, angle of incidence, and the composition of the background atmosphere

In ion milling, the positive ions are generated in a confined plasma discharge and accelerated in the form of a beam towards the sample to be etched, as shown in Fig. 3.31. Inert gases, such as argon, are usually used in these systems because they exhibit a higher sputtering yield than other atoms and also because they do not participate in chemical reactions. A neutralizer, usually a hot filament, emits a flux of electrons to cancel the positively charge of the ions while keeping their kinetic energy for etching. The sample can thus be kept neutral, so that no deleterious charging effects occur. The sample resides in a moderate vacuum (10^{-4}~10^{-6} Torr) environment and can be attached to a cooled substrate holder to maintain its surface at low temperatures during etching. The sample can also be positioned at any angle with respect to the incoming ion beam.

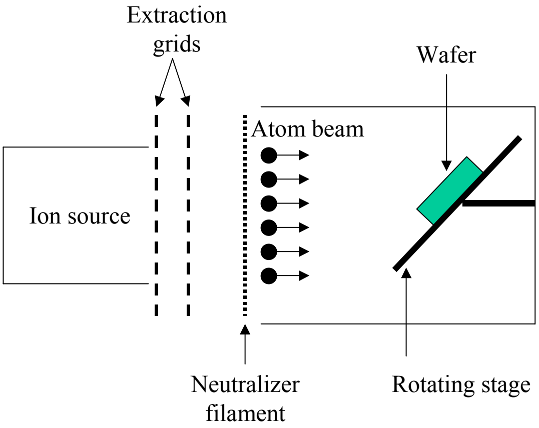


Fig. 3.31. Schematic diagram of an ion-beam milling system. The entire system is under moderate vacuum. The positive ions of inert gas are generated in a confined plasma discharge, away from the wafer, and accelerated in the form of a beam towards the sample to be etched. A neutralizer, usually a hot filament, emits a flux of electrons to cancel the positively charge of the ions. The wafer can be attached to a cooled substrate holder and be positioned at any angle with respect to the incoming ion beam.

Although there are similarities between ion milling and sputter etching, the ion milling process offers more flexibility, can be carried out at a lower temperature, in a less harsh etch environment, and results in a reduced redeposition of contaminants.

The main advantage of ion milling over other methods is the absence of undercutting in this process. However, this technique suffers from a number of disadvantages: it is a slow process, it generates a good amount of heat which makes the subsequent removal of the resist difficult, the sputtered material can redeposit anywhere on the wafer surface, scattering effects make the etching of vertical edges much faster, and trenching effects can occur as a result of the sample tilting.

3.5. Metallization

3.5.1. Metal interconnections

In addition to lithography and etching, a third important step in the fabrication process of a semiconductor device is the deposition of metals in certain areas of the wafer, through a process called metallization. This is done in order to allow the surface wiring of individual semiconductor layers and metal interconnection between contacts in a microcircuit as shown in Fig. 3.32.

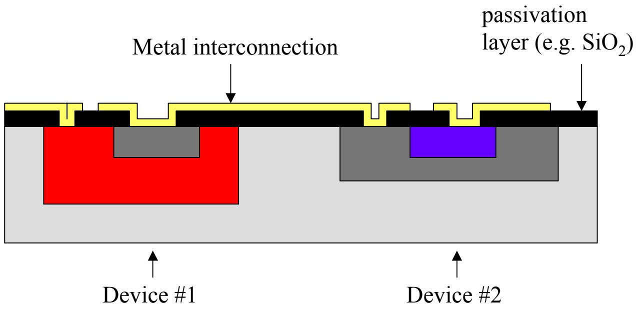


Fig. 3.32. Single layer metal interconnection between two devices. A passivation layer, such as silicon dioxide, is generally first deposited and patterned to isolate areas of the devices which must be electrically connected. Then a metal layer is deposited and patterned to connect electrically one part of a device to another part of a second device.

The metal interconnection is first deposited in the form of thin films of various materials on the surface of the semiconductor wafer. The thicknesses of these films are typically on the order of 1,500 to 15,000 Å. There exist several deposition techniques which will be discussed below. However, in order to be useful, the metal must not cover the entire wafer uniformly, but only certain areas. Lithography is then used to define the areas where the metal will remain. The direct patterning and the lift-off techniques are equally used in this process. A few examples of wet chemical etching solutions for various metals were given in Table 3.2. In order to isolate metal interconnects from one another, to prevent current leakage and short circuits, a protective or passivation layer of dielectric material (e.g. silicon dioxide) is often used, as shown in Fig. 3.32. In most cases, following the formation of the metal interconnects, a heat treatment step called alloying is performed typically between 200 and 400 °C in order to ensure good mechanical and electrical contact between the metal and the semiconductor wafer surface.

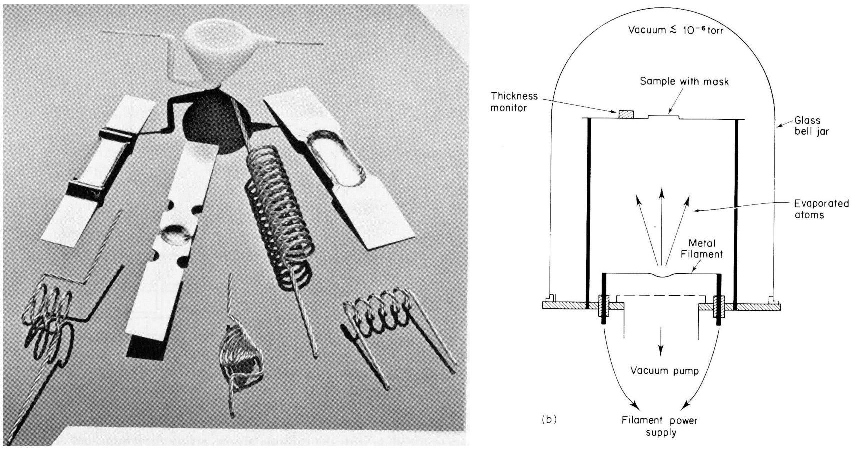
The metal materials used must ideally satisfy a number of properties, such as: have a good electrical current-carrying capability, a good adhesion to the top surface of the wafer, a good electrical contact with the wafer material, be easy to pattern (etch or lift-off), be of high purity, be corrosion resistive, and have long-term stability.

Most of these characteristics are met by either gold or aluminum. In microelectronic circuit technology, aluminum is the most commonly used metal interconnect because it adheres well to both silicon and silicon dioxide, although it is less conductive than copper or gold. Aluminum also has a good current-carrying capability and is easy to pattern with conventional lithography process.

However, one metal material is often not sufficient to satisfy all the properties mentioned previously. This is why most circuit designs require the use of multilayer metal films, such as platinum and gold or a combination of titanium, platinum and gold. A multilayer metal stack makes it possible to avoid the use of gold as a direct electrical contact with silicon because gold adheres badly to the semiconductor surface and is at the origin of significant current leakage which can impair the device performance.

3.5.2. Vacuum evaporation

The deposition of metal thin films on a semiconductor wafer is commonly accomplished through vacuum deposition. Fig. 3.33(a) shows a typical vacuum deposition system. It consists of a vacuum chamber, maintained at a reduced pressure by a pumping system. The shape of the chamber is generally a bell jar that is made of quartz or stainless steel, inside which many components are located, including: the metal sources, a wafer holder, a shutter, a thickness rate monitor, heaters, and an ion gauge to monitor the chamber pressure.



(a) (b)

Fig. 3.33. (a) Examples of evaporation sources which can be used in vacuum evaporation systems; (b) cross-section of a typical vacuum evaporation system which includes a glass bell jar under vacuum, a sample holder, a metal filament (evaporation source), and a thickness monitor. [An Introduction to Semiconductor Microtechnology, Morgan, D.V. and Board, K. Copyright 1990. © John Wiley & Sons Limited. Reproduced with permission.]

It is important to operate at a reduced pressure for a number of reasons. First is a chemical consideration. If any air or oxygen molecule is found in the vacuum chamber during the evaporation of aluminum, the metal would readily oxidize and aluminum oxide would form in the depositing film. Reducing the pressure ensures that the concentration of residual oxygen molecules is small enough to minimize the oxidation reaction. Secondly, the coating uniformity is enhanced at a higher vacuum. Indeed, at low pressures, the mean free path of the evaporated metal atoms is increased, that is the distance traveled by the atoms before collision with another atom. When the mean free path exceeds the dimensions of the chamber, this ensures that the metal atoms will strike the wafers before hitting another atom which would have caused non-uniform depositions.

Two ranges of vacuum conditions are typically used for vacuum evaporation: the low and medium vacuum range ($10^5 \sim 10^{-1}$ Pa) and the high vacuum range ($10^{-1} \sim 10^{-4}$ Pa). For the low-medium vacuum range, a mechanical roughing pump is sufficient. To attain the high vacuum range, a roughing pump is first used to evacuate the chamber to the medium vacuum range, then a high capacity pump takes over. The most common high capacity vacuum pumps include diffusion pumps, cryo pumps, and turbomolecular pumps.

The physical laws governing the evaporation of metal are those of the kinetic theory of gases in which each particle, e.g. metal atom or gas molecule, is modeled as moving freely in space with a momentum and energy, which is subject to instantaneous collision events with other particles, the probability for a collision to occur is proportional to the interval of time since the previous collision, and the particles reach thermal equilibrium only through such collisions. In this model, the mean free path of a particle is given by:

Eq. (3.4) $$\lambda = \frac{k_b T}{\sqrt{2}\pi P d^2}$$

where k_b is the Boltzmann constant, T the absolute temperature, d is the diameter of the particle, and P is its partial pressure in the chamber which is related to the number of particles in the chamber N through the ideal gas law:

Eq. (3.5) $$N = \frac{PV}{k_b T}$$

where V is the volume of the chamber. A high quality film can only be obtained with a clean environment, e.g. a clean chamber, pure source material and clean wafer surface.

There are three different types of evaporation techniques, depending on the method used to physically evaporate the metal from its solid state: filament evaporation, electron-beam evaporation, and flash hot plate.

The filament evaporation is the simplest of these methods. The metal can be in the form of a wire wrapped around a coiled tungsten that can sustain high temperatures and current as shown Fig. 3.33(b). The metal can also be stored in tungsten boats if large quantities of material are required. An electrical current is passed through the tungsten boat, thus heating and melting the metal into a liquid which can then evaporate into the chamber at low pressure. Filament evaporation is not very controllable due to temperature variations along the filament. Another drawback when using filaments is that the source material can easily be contaminated and the contaminants can subsequently be evaporated onto the wafers. Moreover, mixtures of metal alloys containing for example titanium, platinum, nickel, and gold are difficult to achieve using the filament evaporation method because each metal has a different evaporation rate at a given temperature.

To avoid such problems, the electron-beam evaporation technique was developed. Fig. 3.34 is an illustration of the principle of an electron-beam source, which consists of a copper holder or crucible with a center cavity which contains the metal material. A beam of electrons is generated and bent by a magnet flux so that it strikes the center of the charge cavity as shown in Fig. 3.34. In addition, the solid metal within the crucible is heated to its melting point such that it presents a smooth and uniform surface where the electron beam hits, thus ensuring that the deposition on the wafer is uniform. The crucible is cooled with water to maintain the edges of the metal in a solid state. Electron-beam evaporation is relatively controlled for a variety of metals such as aluminum and gold. This method had its own limitations: it can only evaporate one alloy at a time. But, over the years electron-beam systems have incorporated multiple guns so that each material will have its own electron beam.

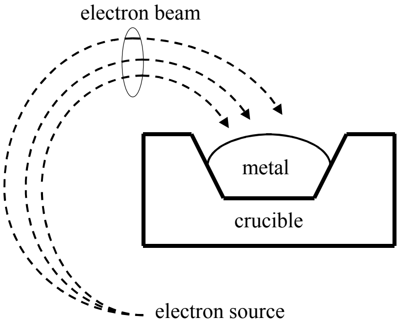


Fig. 3.34. Schematic diagram of an electron-beam evaporation source in which electrons are generated by a high temperature filament and accelerated into an electron beam. A magnetic field bends the electron beam so that it hits a metal charge in a water-cooled crucible. The impact melts the metal and allows it to evaporate in the vacuum of the deposition chamber.

The *flash hot plate* method uses a fine wire as the source material. This fine wire which contains an alloy material is fed automatically onto a hot plate surface. Upon contact the tip of the wire melts and the material "flashes" into a vapor and coats the wafers in the chamber. Since all of the elements are flashed simultaneously, the composition of the metal film deposited on the wafer is close to the alloy composition of the wire.

3.5.3. Sputtering deposition

Sputtering deposition is also called physical vapor deposition and is a physical process. A typical sputtering deposition system is shown in Fig. 3.35. It contains a slab or target of the desired metal which is electrically grounded and serves as the cathode. Under vacuum conditions, argon gas is introduced into the chamber and is ionized into a positively charged ion. These are accelerated toward the cathode target. By impacting the target, enough metal atoms are dispersed such that they deposit onto the wafer surface. The main feature of the sputtering method is that the target material is deposited on the wafer without chemical or compositional change.

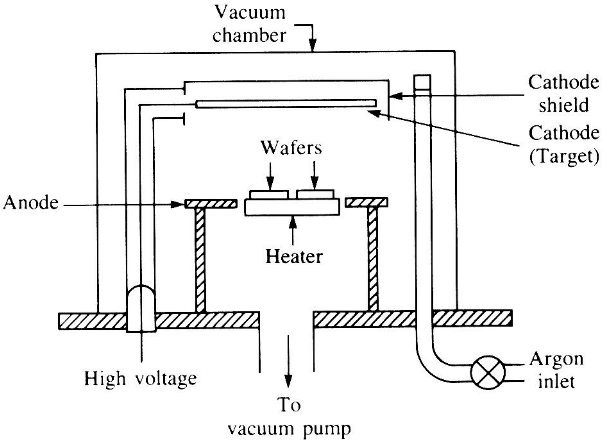


Fig. 3.35. Cross-section schematic diagram of a typical sputtering system, which is enclosed in a vacuum chamber and includes the wafers which are placed on a heater, and a set of electrodes, one of which is made from the target material to be sputtered. Argon gas is supplied and ionized so that ions can impact on the target to release atoms of the material to be deposited. [Jaeger, R.C., Introduction to Microelectronics Fabrication: Vol. 5 of Modular Series on Solidstate Devices, 2nd Edition,p. 114, fig. 6.6 © 2002. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.]

Sputtering has several advantages over other traditional evaporation techniques. For example, the composition of the deposited film is precisely determined by that of the target material, step coverage is improved, and sputtered films have a higher adhesion.

As with the evaporation technique, a high quality film can only be obtained with a clean environment, e.g. clean chamber, pure source material and clean wafer surface.

3.6. Packaging of devices

The final step in the fabrication of a semiconductor device consists of separating the individual components on a same wafer and packaging them.

3.6.1. Dicing

At the industrial scale, mass produced wafers contain a large number of equivalent integrated circuits which need to be separated from one another. Each resulting circuit is called a die chip. This can be accomplished for example by using a diamond saw as shown in Fig. 3.36 (a). As the demand for accuracy becomes important and tolerances get tighter, other forms of separation have been developed including for example a laser beam as shown in Fig. 3.36(c).

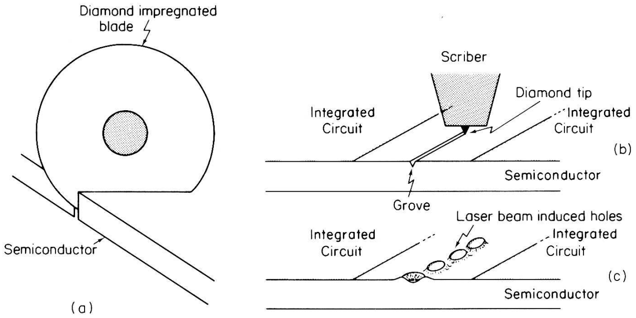
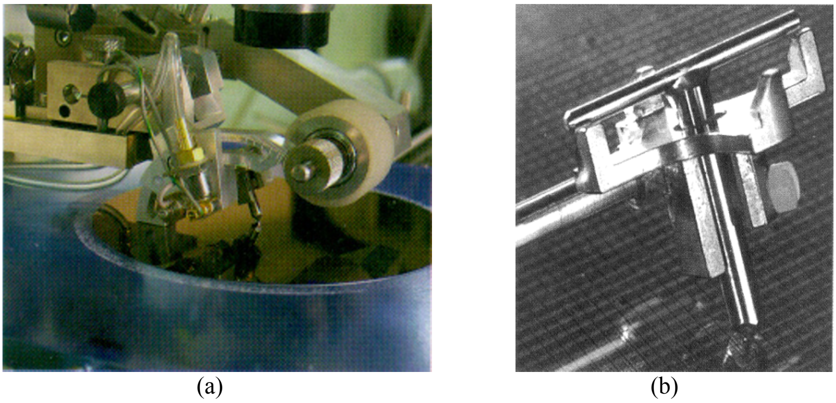

Diamond impregnated blade
Semiconductor
(a)
Scriber
Diamond tip
Integrated Circuit
Integrated Circuit
(b)
Semiconductor
Grove
Laser beam induced holes
Integrated Circuit
Integrated Circuit
(c)
Semiconductor

Fig. 3.36. Illustration of the various methods that can be used to separate individual devices from a semiconductor wafer by using: (a) a diamond saw, (b) a scriber, or (c) a laser beam. [An Introduction to Semiconductor Microtechnology, Morgan, D.V. and Board, K. Copyright 1990. © John Wiley & Sons Limited. Reproduced with permission.]

Fig. 3.37(a) is a photograph of a modern commercial scribing tool, while Fig. 3.37(b) is a close up look at the scriber tip.



(a) (b)

Fig. 3.37. (a) Photograph of a modern commercial scribing tool showing the scriber tip, with its positioning wheels, and a camera. (b) Close up photograph of the scriber tip. [Reprinted with permission from Loomis Industries, Inc.]

Fig. 3.38 is a photograph illustrating a wafer after scribing on which one can see delimitated chip-scale die components. Following the dicing of the wafer, each individual die chip is sorted and inspected under a microscope before wire bonding and packaging.

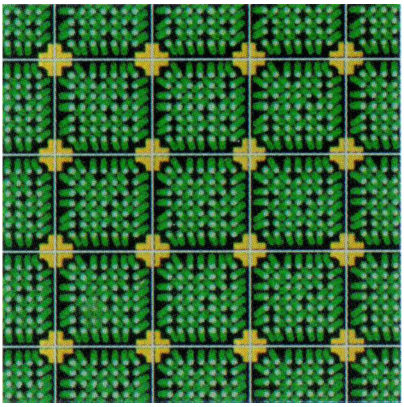


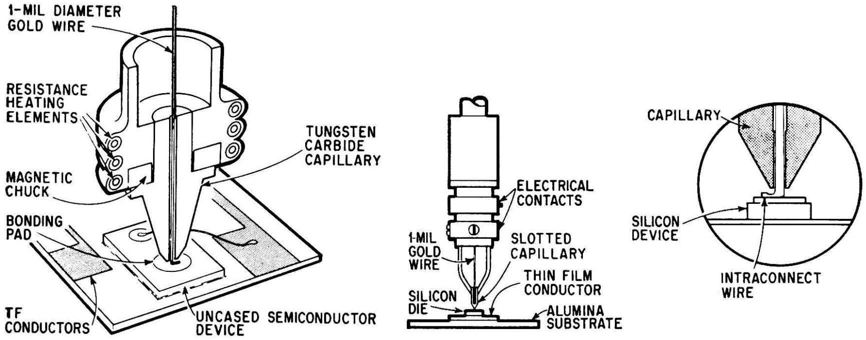
Fig. 3.38. Photograph of chip-scale die components delimitated on a wafer after scribing. [Reprinted with permission from Kulicke & Soffa Industries.]

3.6.2. Wire bonding

It is necessary to link the metal interconnects which have microscopic sizes to a macroscopic electrical connector. The method used is called wire bonding. The wire used consists of gold or aluminum with a diameter of about 10 to 50 micrometers. Gold wire is generally used in industry as it welds readily to both aluminum and gold contact pads by heat and pressure. This process is known also as thermocompression bonding.

A fine wire of gold is fed through a resistance-heated tungsten carbide capillary tube as shown in Fig. 3.39(a). Applying an electric spark melts the exposed end of the wire which is brought down with pressure upon the area of the metal contact where it is then welded. Under manual or automated control, the capillary is moved to another contact pad where the second bond will be made. The capillary is then raised and the wire is broken near the edge of the bond by an electric spark which forms a ball.

A variation of this technique is the pulse-heated thermocompression bonding method, as shown in Fig. 3.39(b), in which the tungsten carbide bonding tool is heated by a pulse of current rather than an electric resistance.


1-MIL DIAMETER GOLD WIRE
RESISTANCE HEATING ELEMENTS
MAGNETIC CHUCK
BONDING PAD
TF CONDUCTORS
UNCASED SEMICONDUCTOR DEVICE
TUNGSTEN CARBIDE CAPILLARY
ELECTRICAL CONTACTS
1-MIL GOLD WIRE
SLOTTED CAPILLARY
THIN FILM CONDUCTOR
SILICON DIE
ALUMINA SUBSTRATE
CAPILLARY
SILICON DEVICE
INTRACONNECT WIRE

(a) (b)

Fig. 3.39. Schematic diagrams of (a) a resistance-heated thermocompression wire bonder, and (b) a pulse-heated thermocompression wire bonding tool. Applying an electric spark melts the exposed end of the wire which is brought down with pressure upon the area of the metal contact where it is then welded. Under manual or automated control, the capillary is moved to another contact pad where the second bond will be made. [Fogiel, M., Microelectronics-Principle, Design Techniques, and Fabrication Processes. Copyright © 1968 by Research & Education, Inc. Reprinted by permission of Research and Education Association, New York.]

The sequence of steps during these thermocompression processes is schematically illustrated in Fig. 3.40. An example of wire-bonded die viewed at high magnification under electron microscopy is shown in Fig. 3.41.

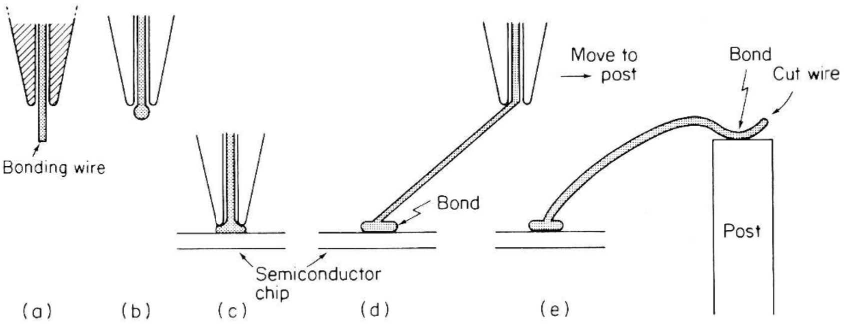


Fig. 3.40. Schematic diagrams showing the sequence in the thermocompression wire bonding process. [An Introduction to Semiconductor Microtechnology, Morgan, D.V. and Board, K. Copyright 1990. © John Wiley & Sons Limited. Reproduced with permission.]

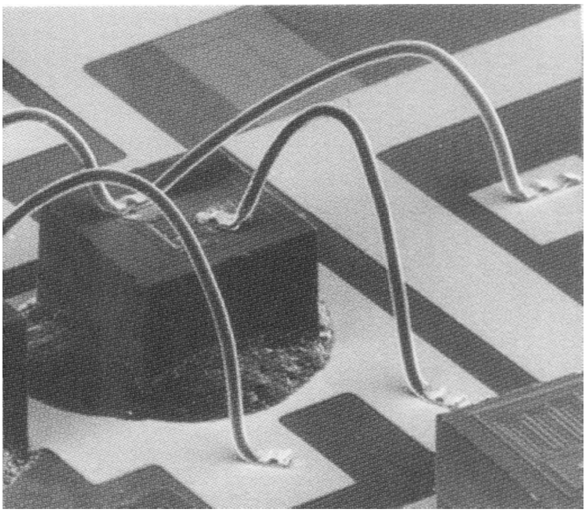


Fig. 3.41. An example of wire-bonded die viewed under electron microscopy. [Reprinted with permission from Kulicke & Soffa Industries.]

While heating of the wire works quite well for gold, when aluminum wire is used the high temperatures that are required cause oxidation which makes it difficult to form a good ball at the end of the wire. Thus an alternative process is needed, it is known as ultrasonic bonding. In this technique the bond is formed by pressure and mechanical vibration. In this case as the wire leaves the spool, the tip is pushed against the surface and the vibration removes any existing oxide and allows the metal to deform and flow under pressure at room temperature to create a strong bond. The result is a good bond with little to now oxide formation.

3.6.3. Packaging

Once the die chip is fully wire bonded, it is ready to be encapsulated in a package. Integrated circuit devices can be mounted in a wide variety of packages which have a specialized shape and nature. In this sub-section, we will briefly review three examples of packages shown in Fig. 3.42.

Fig. 3.42(a) shows a round TO-style package which is commonly used for low power transistors. The package utilizes a pie shape header where the silicon chip or die is mounted to the center of the gold plated header. Wires are connected from the die pad to the Kovar lead posts that protrude through the header. A glass-to-metal cap is sealed over the die chip to protect the device.

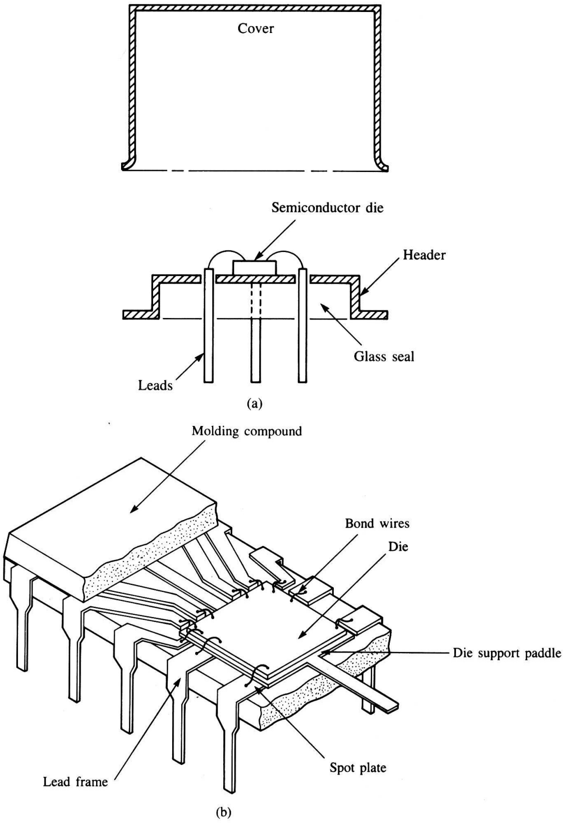

Cover
Semiconductor die
Header
Glass seal
Leads
(a)
Molding compound
Bond wires
Die
Die support paddle
Lead frame
Spot plate
(b)

Fig. 3.42. Schematic diagram of (a) a TO-style package and (b) a dual line package. [Reprinted with permission from 19th IEEE International Reliability Physics Symposium Proceedings, Howell, J.R., "Reliability study of plastic encapsulated copper lead frame epoxy die attach packaging system," pp. 104-110. © 1981 IEEE.]

Another form of packages is the dual line package (DIP) as illustrated in Fig. 3.42(b). The dual line package is considered the least expensive package and the most popular one in industry. The design of the DIP package is such that it eliminates the waste of volume of the TO can, and brings the die closer to the metal leads. One advantage of the dual line package over TO packages is the amount of leads that can pass through the walls. Typically, dual line packages contain four to eighty leads. Leads projecting from the walls of the package rather than at the base can bring out more leads from a package of a given size, and still maintain reasonable space between the leads.

Both “TO”-style packages and dual-in-line packages are packages designed for surface mounting; that is they are both designed to be mounted into prepatterned holes on printed circuit boards (PCBs). While these types of mounts are used for making most systems, they require PCBs to be made before any testing can be performed; this is quite expensive. A method that was developed to permit processing of batch fabricated controls electronics with the desired circuit is known as flip-chip bonding. In this method the control electronics and active system are fabricated separately then sandwiched together to form both the package and interconnection. A schematic diagram of a hybridized focal plane array is shown in Fig. 3.43.

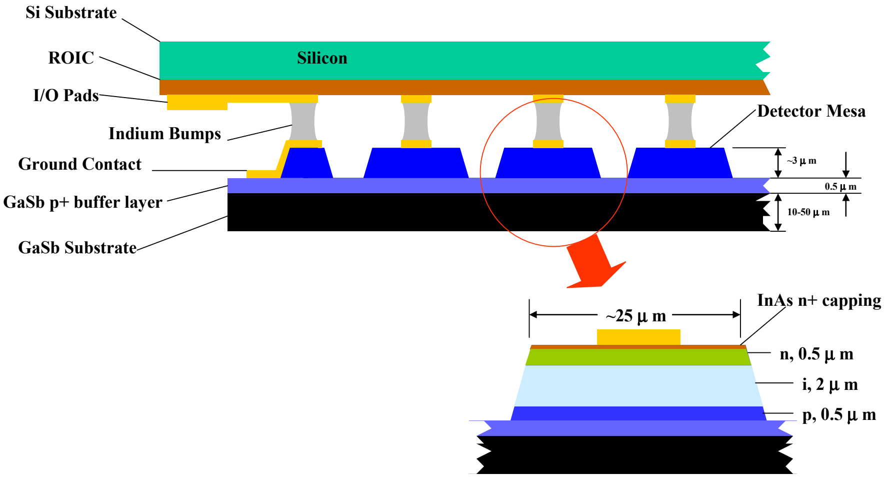


Fig. 3.43. In this system the Si-based control electronics and the p-i-n photodetectors are formed separately. They both have indium solder ball formed on their contacts. They are then aligned, the temperature of the device is then heated, to allow the solder to reflow and create both the electrical contact and the die connection to occur simultaneously.

3.7. Summary

In this Chapter, we have reviewed the important steps involved in the fabrication of a semiconductor device. We described the photolithography and the electron-beam lithography processes. We showed the difference between positive and negative resists, and between the direct patterning and lift-off techniques. We have discussed the various etching process which are commonly used, including wet chemical etching, plasma etching, reactive ion etching, sputter etching and ion milling. We described the metallization process, including the deposition of metal thin films and the formation of metal interconnections. Finally, we presented in broad lines the packaging of semiconductor devices, which involves the dicing of the wafer into chip dies, their wire bonding and packaging into standard packages.

References

Campbell, S.A., *The Science and Engineering of Microelectronic Fabrication*, Oxford University Press, New York, 1996.

Fogiel, M., *Microelectronics-Principle, Design Techniques, and Fabrication Processes*, Research and Education Association, New York, 1968.

Howell, J.R., “Reliability study of plastic encapsulated copper lead frame epoxy die attach packaging system,” *19th IEEE International Reliability Physics Symposium Proceedings*, pp. 104-110, 1981.

Jaeger, R.C., *Introduction to Microelectronic Fabrication*, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ, 2002.

Morgan, D.V. and Board, K., *An Introduction to Semiconductor Microtechnology*, John Wiley & Sons, Chichester, UK, 1990.

Piner, R.D., Zhu, J., Xu, F., Hong, S., and Mirkin, C.A., “Dip-pen nanolithography,” *Science* 283, pp. 661-663, 1999.

Wu, W., Cui, B., Sun, X.Y., Zhang, W., Zhuang, L., Kong, L., and Chou, S.Y., “Large area high density quantized magnetic disks fabricated using nanoimprint lithography,” *Journal of Vacuum Science and Technology* 16, pp. 3825-3829, 1998.

Further reading

Choudhury, P.R., *Handbook of microlithography, micromachining & microfabrication*, vol. 1: Microlithography, SPIE Optical Engineering Press, Bellingham, WA, 1997.

Castaño, J.L., Piqueras, J., Gomez, L.J., and Montojo, M.T., “Chemical cleaning of GaSb (1,0,0) surfaces,” *Journal of the Electrochemical Society* 136, pp.1480-1484, 1989.

D'Agostino, R., Cramarossa, F., Fracassi, F., Desimoni, E., Sabbatini, L., Zambonin, P.G., and Caporiccio, G., "Polymer film formation in C_2F_6-H_2 discharges," *Thin Solid Films* 143, pp. 163-175, 1986.

Diaz, J.E., *Fabrication of High Power Aluminum-Free 0.8 μm to 1.0 μm InGaP/InGaAsP/GaAs Lasers for Optical Pumping*, PhD dissertation, Northwestern University, 1997.

Elliott, D.J., *Microlithography: Process Technology for IC Fabrication*, McGraw-Hill, New York, 1986.

Elliott, D.J., *Integrated Circuit Fabrication Technology*, 2nd edition, McGraw-Hill, New York, 1989.

Ghandi, S., *VLSI Fabrication Principles*, John Wiley & Sons, New York, 1983.

Hatzakis, M., "High sensitivity resist system for lift-off metallization," U.S. Patent No. 4024293, 1975.

Levinson, H.J., *Principles of Lithography*, SPIE Optical Engineering Press, Bellingham, WA., 2001.

Mackie, S. and Beaumont, S.P., "High sensitivity positive electron resist," *Solid State Technology* 28, pp. 117-122, 1985.

Madou, M.J., *Fundamental of Microfabrication*, CRC Press, Boca Raton, FL, 1997.

Moreau, W. and Ting, C.H., "High sensitivity positive electron resist," US Patent 3934057, 1976.

Plummer, J.D., Deal, M., and Griffin, P.B., *Silicon VLSI Technology: Fundamentals, Practice and Modeling*, Prentice-Hall, Upper Saddle River, NJ, 2000.

Propst, E.K., Vogt, K.W., and Kohl, P.A., "Photoelectrochemical Etching of GaSb," *Journal of the Electrochemical Society* 140, pp. 3631-3635, 1993.

Sheats, J.R. and Smith, B.W., *Microlithography: Science and Technology*, Marcel Decker, New York, 1998.

Williams, R.E., *Gallium Arsenide Processing Technique*, Artech House, Dedham, MA, 1984.

Wong, A.K., *Resolution Enhancement Techniques in Optical Lithography*, SPIE Optical Engineering Press, Bellingham, WA, 2001.

Problems

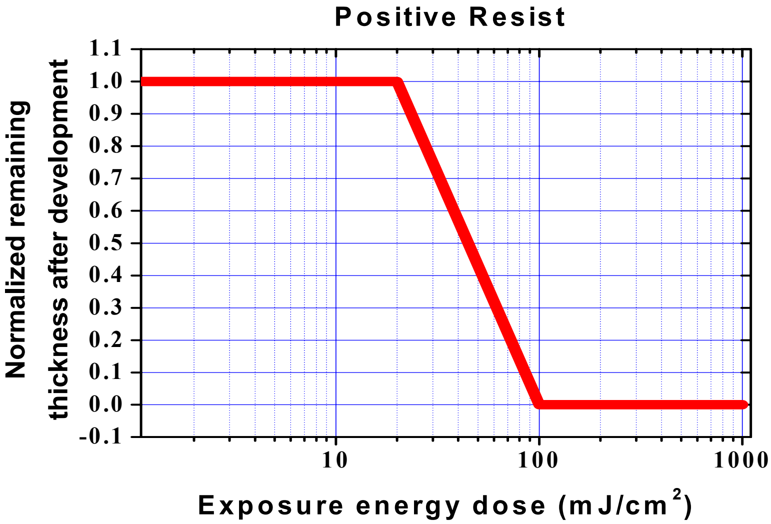
1. Draw a layout, in top view and perspective, where the mask would be opaque to realize an "+" shaped metal line using a positive resist and the lift-off technique.

2. Do the same as Problem 1, but using a negative resist.

3. Design a photolithography and metallization sequence of steps to obtain a Au square and a Ti circle on the surface of a semiconductor wafer. Draw the shape of the mask to be used. At each appropriate step, indicate whether you use the direct patterning or the lift-off technique, you use positive or negative resist.

4. Indicate the advantages and disadvantages of using gold or aluminum for wire bonding. Which metal has a higher electrical conductivity?

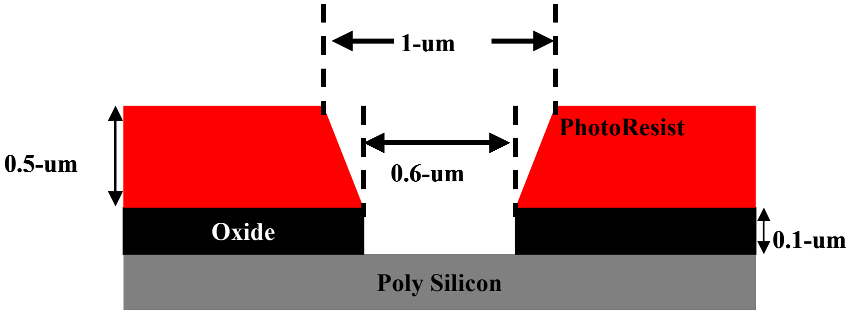
5. For a particular positive resist, the normalized remaining thickness after development versus photoexposure energy density is plotted below. Calculate the contrast (γ) of this resist ($\gamma = 1/\log(E_T/E_1)$).



6. Draw a layout, in side view of a set of laser bars (i.e. a comb structure) if you were using a isotropic etch and if you were using an anisotropic etchant. Which etchant gives a higher aspect ratio?

7. Discuss the major advantages and disadvantages of electron beam lithography.

8. Poly-Si of the following structure is to be etched using a *completely anisotropic* dry-etch process, to remove poly-Si at a rate of 0.1 µm /min. However, this etch process has poor selectivities: selectivity to SiO_2 is 5:1, selectivity to photoresist is 2:1.
   (a) Sketch the cross-section after 5 minutes of etching.
   (b) Calculate the angle of the SiO_2 sidewalls after 5 minutes of etching.



9. In this Chapter we have discussed several different forms of lithography (i.e. photo, electron beam, x-ray, high energy ion beam). Why do companies like Intel keep developing new lithographic techniques?

10. What is the wavelength regime of each of the lithographic techniques listed in Problem 9? (red-visible light, green-visible light, blue light, UV (E = 5 eV), e-beam (acceleration voltage of 10 keV), and x-ray).

11. Compare wet and dry etching in terms of its directionality, selectivity, cleanliness, feature size, and controllability.

4. Semiconductor *p-n* and Metal-Semiconductor Junctions

4.1. Introduction
4.2. Ideal *p-n* junction at equilibrium
  4.2.1. Ideal *p-n* junction
  4.2.2. Depletion approximation
  4.2.3. Built-in electric field
  4.2.4. Built-in potential
  4.2.5. Depletion width
  4.2.6. Energy band profile and Fermi energy
4.3. Non-equilibrium properties of *p-n* junctions
  4.3.1. Forward bias: a qualitative description
  4.3.2. Reverse bias: a qualitative description
  4.3.3. A quantitative description
  4.3.4. Depletion layer capacitance
  4.3.5. Ideal *p-n* junction diode equation
  4.3.6. Minority and majority carrier currents in neutral regions
4.4. Metal-semiconductor junctions
  4.4.1. Formalism
  4.4.2. Schottky and ohmic contacts
4.5. Summary

4.1. Introduction

Until now, our discussion was based solely on homogeneous semiconductors whose properties are uniform in space. Although a few devices can be made from such semiconductors, the majority of devices and the most important ones utilize non-homogeneous semiconductor structures. Most of them involve semiconductor *p-n* junctions, in which a *p*-type doped region and an *n*-type doped region are brought into contact. Such a junction actually forms an electrical diode. This is why it is usual to talk about a *p-n* junction as a diode. Another important structure involves a semiconductor in intimate contact with a metal, leading to what is called a metal-

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_4,
© Springer Science+Business Media, LLC 2010

semiconductor junction. Under certain circumstances, this configuration can also lead to an electrical diode.

The objective of this Chapter will first be to establish an accurate model for the *p-n* junction which can be at the same time mathematically described. This model will be the ideal *p-n* junction diode. The basic properties of this ideal *p-n* junction at equilibrium will be described in detail. The non-equilibrium properties of this *p-n* junction will then be discussed by deriving the diode equation which relates the current and voltage across the diode. Deviations from the ideal diode case will also be described. Finally, this Chapter will also discuss the properties of metal-semiconductor junctions and compare them with those of *p-n* junctions.

4.2. Ideal *p-n* junction at equilibrium

4.2.1. Ideal p-n junction

The ideal *p-n* junction model is also called the abrupt junction or step junction model. This is an idealized model for which we assume that the material is uniformly doped *p*-type with a total acceptor concentration N_A on one side of the junction (e.g. $x < 0$), and the material is uniformly doped *n*-type with a total donor concentration N_D on the other side (e.g. $x > 0$). For further simplicity, we will consider a homojunction, i.e. both doped regions are of the same semiconductor material. We will restrict our analysis to the one-dimensional case, as illustrated in Fig. 4.1.

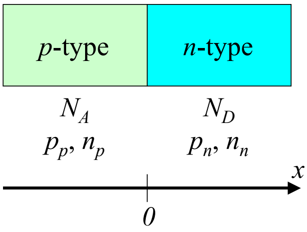


Fig. 4.1. Ideal p-n junction model, in which one side of the junction is a purely p-type semiconductor and the other a purely n-type semiconductor. Both materials are uniformly doped.

In the *p*-type doped region far from the junction area, the equilibrium hole and electron concentrations are denoted p_p and n_p, respectively. In the *n*-type doped region far from the junction area, the hole and electron concentrations are denoted p_n and n_n, respectively. These carrier concentrations satisfy the mass action law $\left(n_i^2 = np\right)$.

Eq. (4.1) $$p_p n_p = p_n n_n = n_i^2$$

where n_i is the intrinsic carrier concentration in the semiconductor material considered. We further assume that all the dopants are ionized, which leads to the following carrier concentrations for the *p*- and *n*-type regions, respectively:

Eq. (4.2) $$\begin{cases} p_p = N_A & \left(10^{16}\, cm^{-3}\right) \\ n_p = \dfrac{n_i^2}{N_A} & \left(10^{5}\, cm^{-3}\right) \end{cases} \text{ and } \begin{cases} n_n = N_D & \left(10^{17}\, cm^{-3}\right) \\ p_n = \dfrac{n_i^2}{N_D} & \left(10^{4}\, cm^{-3}\right) \end{cases}$$

A few typical values for these concentrations are given in parenthesis. It is important to remember that both a *p*-type, and an *n*-type, isolated semiconductors are electrically neutral.

4.2.2. Depletion approximation

However, when bringing a *p*-type semiconductor into contact with an *n*-type semiconductor, the material is not electrically neutral everywhere anymore. Indeed, on one side of the junction area, for $x < 0$, there is a high concentration of holes whereas on the other side there is a low concentration of holes. This asymmetry in carrier density results in the diffusion of holes across the junction as shown in Fig. 4.2. By doing so, the holes leave behind uncompensated acceptors ($x < 0$) which are negatively charged. A similar analysis can be carried out for electrons as there is also a asymmetry in the density of electrons on either side of the *p-n* junction. This leads to their diffusion and makes the material positively charged for $x > 0$ as the electrons leave behind uncompensated donors, as shown in Fig. 4.2.

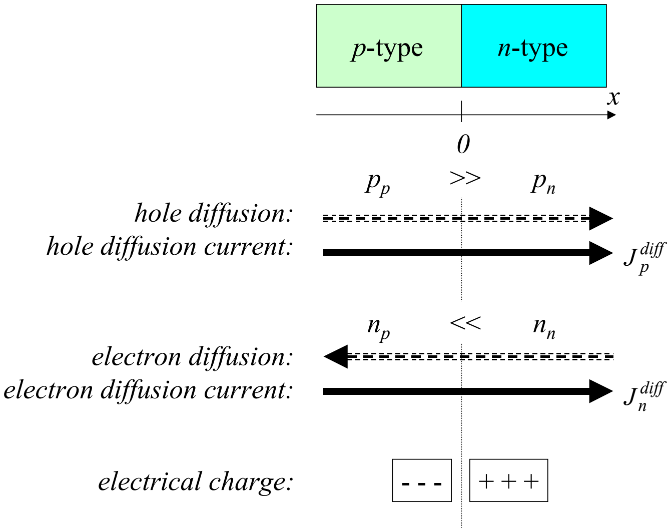


Fig. 4.2. Hole and electron diffusion across a p-n junction. The holes diffuse from the left to the right, which leads to a diffusion electrical current from the left to the right as well. By contrast, the electrons diffuse from the right to the left, but this leads to a diffusion electrical current from the left to the right because of the negative charge of electrons. The diffusion process leaves uncompensated acceptors in the p-type region and donors in the n-type regions, i.e. a net negative charge in the p-type region and a net positive charge in the n-type region. The presence of these charges result in a built-in electric field.

This redistribution of electrical charge does not endure indefinitely. Indeed, as positive and negative charges appear on the $x > 0$ and $x < 0$ sides of the junction respectively, an electric field strength *E(x)*, called the built-in electric field, will result and is shown in Fig. 4.3. This electric field will generate the drift of the positively charged holes and the negatively charged electrons. By comparing Fig. 4.2 and Fig. 4.3, we can see that the drift of these charge carriers counteracts the previous diffusion process. An equilibrium state is reached when the diffusion currents $J^{diffusion}$ and drift currents J^{drift} are exactly balanced for each type of carrier, i.e. holes and electrons taken independently:

Eq. (4.3)
$$\begin{cases} J_h^{diff} + J_h^{drift} = 0 \\ J_e^{diff} + J_e^{drift} = 0 \end{cases}$$

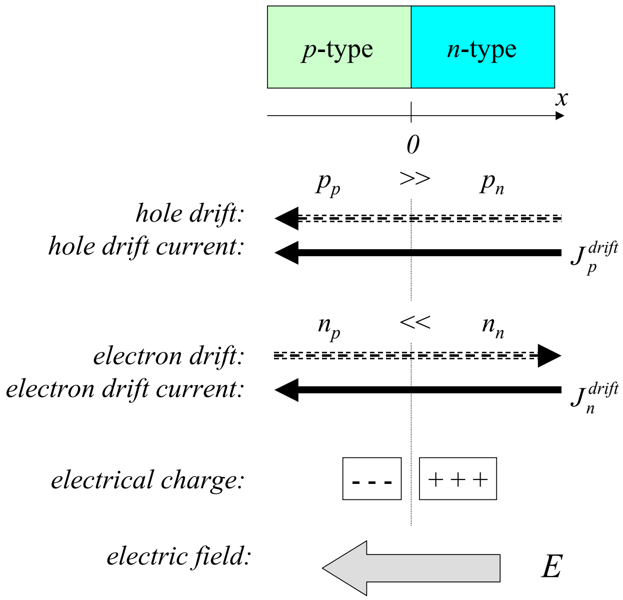


Fig. 4.3. Hole and electron drift across a p-n junction. Under the influence of the built-in electric field, the holes drift from the right to the left, which leads to a drift electrical current from the right to the left as well. By contrast, the electrons drift from the left to the right, but this leads to a drift electrical current from the right to the left because of the negative charge of electrons. The drift process counterbalances the diffusion of charge carriers in order to bring the system into equilibrium.

There is a transition region around the *p-n* junction area with a width W_0 in which the electrical charges are present. This region is called the space charge region and is schematically shown in Fig. 4.4(a). The charge distribution within this region is modeled as follows: we consider that there is a uniform concentration of negative charges for $-x_{p0} < x < 0$ equal to $Q(x) = -qN_A$ (where N_A is the total concentration of acceptors in the *p*-type region), and a uniform concentration of positive charges for $0 < x < x_{n0}$ and equal to $Q(x) = +qN_D$ (where N_D is the total concentration of donors in the *n*-type region). The quantities x_{p0} and x_{n0} are positive and express how much the space charge region extends on each side of the junction, as illustrated in Fig. 4.4(b). The width of the space charge region, also called depletion width, is then given by:

Eq. (4.4) $$W_0 = x_{n0} + x_{p0}$$

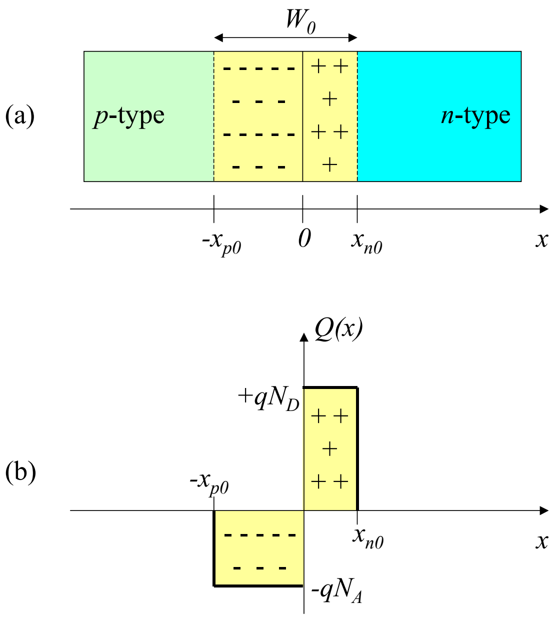


Fig. 4.4. (a) Space charge region in a p-n junction. Near the junction area, the p-type region is negatively charged as a result of the diffusion of charge carriers. (b) Electrical charge density in a p-n junction. To keep the overall charge neutrality, the total number of negative charges in the p-type region is equal to the total number of positive charges in the n-type region. In the depletion approximation, the charges are assumed uniformly distributed in space, within the depletion region delimited by $-x_{p0}$ *and* x_{n0}.

Outside of this space charge region, we assume that the semiconductor is electrically neutral without any charge depletion and that the hole and electron concentrations are given by Eq. (4.2). These regions will be called the bulk p-type and bulk n-type region. The carrier concentrations must therefore somehow go from a high value on one side of the junction to a low value on the other side, and this occurs within the space charge region, as illustrated in Fig. 4.5. In particular, we have:

Eq. (4.5)

$$\begin{cases} p(-x_{p0}) = p_p \quad \text{and} \quad p(x_{n0}) = p_n \\ n(-x_{p0}) = n_p \quad \text{and} \quad n(x_{n0}) = n_n \end{cases}$$

This model is called the depletion approximation. In this model, there are no free holes or electrons in the space charge region: the depletion of carriers is complete. The electric field exists only within this space charge region.

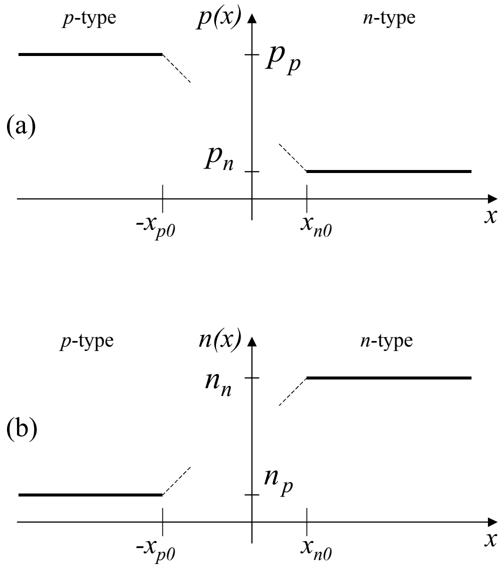


Fig. 4.5. (a) Hole and (b) electron concentrations in a p-n junction. In the depletion approximation, the hole and electron concentrations are assumed to be constant and equal to their equilibrium values outside of the depletion region.

Because the entire *p-n* structure must globally remain electrically neutral, and therefore the space charge region must be neutral as a whole, we must equate the total number of negative charges on one side of the junction to the total number of positive charges on the other side, i.e.:

Eq. (4.6) $$qAN_A x_{p0} = qAN_D x_{n0}$$

where A is the cross-section area of the junction, and after simplification:

Eq. (4.7) $$N_A x_{p0} = N_D x_{n0}$$

Combining Eq. (4.4) and Eq. (4.7), we can express the quantities x_{p0} and x_{n0} as a function of the depletion width W_0:

Eq. (4.8) $$\begin{cases} x_{p0} = \dfrac{N_D}{N_A + N_D} W_0 \\ x_{n0} = \dfrac{N_A}{N_A + N_D} W_0 \end{cases}$$

These show that the space charge region extends more in the *p*-type region than in the *n*-type region when $N_D>N_A$ and reciprocally.

4.2.3. Built-in electric field

The built-in electric field strength can be calculated using Gauss's law which can be written in our one-dimensional model as:

Eq. (4.9) $$\frac{dE(x)}{dx}=\frac{Q(x)}{\varepsilon}$$

where ε is the permittivity of the semiconductor material and $Q(x)$ is the total charge concentration. This relation can be rewritten for either sides of the junction:

Eq. (4.10) $$\begin{cases}\dfrac{dE(x)}{dx}=-\dfrac{qN_A}{\varepsilon} & for\ -x_{p0}<x<0\\ \dfrac{dE(x)}{dx}=\dfrac{qN_D}{\varepsilon} & for\ 0<x<x_{n0}\end{cases}$$

From these relations we see that the electric field strength varies linearly on either side of the junction. By integrating Eq. (4.10) using the boundary conditions assumed in the depletion approximation:

Eq. (4.11) $$E(-x_{p0})=E(x_{n0})=0$$

that the electric field strength is equal to zero at the limits of the space charge region ($x=-x_{p0}$ and $x=x_{n0}$), we obtain successively:

Eq. (4.12) $$\begin{cases}E(x)=\displaystyle\int_{-x_{p0}}^{x}dEdx=\int_{-x_{p0}}^{x}-\frac{qN_A}{\varepsilon}dx & for\ -x_{p0}<x<0\\ E(x)=\displaystyle\int_{x_{n0}}^{x}dEdx=\int_{x_{n0}}^{x}\frac{qN_D}{\varepsilon}dx & for\ 0<x<x_{n0}\end{cases}$$

Eq. (4.13) $$\begin{cases}E(x)=-\dfrac{qN_A}{\varepsilon}(x+x_{p0}) & for\ -x_{p0}<x<0\\ E(x)=\dfrac{qN_D}{\varepsilon}(x-x_{n0}) & for\ 0<x<x_{n0}\end{cases}$$

For $x = 0$, we obtain two expressions for the electric field strength from the two previous expressions for *E(x)*:

Eq. (4.14) $$\begin{cases} E(0) = -\dfrac{qN_A}{\varepsilon}\left(x_{p0}\right) \\ E(0) = \dfrac{qN_D}{\varepsilon}\left(-x_{n0}\right) \end{cases}$$

And these expressions are equal, according to Eq. (4.7). Therefore, the global electrical neutrality of the *p-n* structure ensures the continuity of the built-in electric field strength. A plot of *E(x)* is shown in Fig. 4.6.

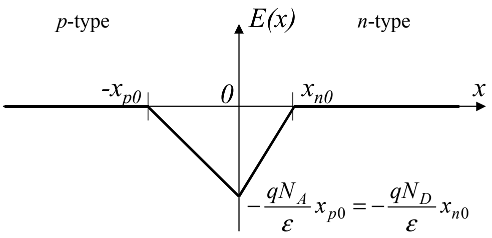


Fig. 4.6. Built-in electric field strength profile across a p-n junction. In the depletion approximation, the electric field strength is zero outside the depletion region because there is no net electrical charge. Within the depletion region, the electric field strength varies linearly with distance.

4.2.4. Built-in potential

As a result of the presence of an electric field, an electrical potential *V(x)* also exists and is related to the electric field strength through:

Eq. (4.15) $$E(x) = -\frac{dV(x)}{dx}$$

The potential is constant outside the space charge region because the electric field strength is equal to zero there. An analytical expression for the electrical potential can be obtained by integrating Eq. (4.13):

Eq. (4.16) $$\begin{cases} V(x) = \dfrac{qN_A}{\varepsilon}\left(\dfrac{x^2}{2} + x_{p0}x\right) & \text{for } -x_{p0} < x < 0 \\ V(x) = -\dfrac{qN_D}{\varepsilon}\left(\dfrac{x^2}{2} - x_{n0}x\right) & \text{for } 0 < x < x_{n0} \end{cases}$$

where we chose the origin of the potential at $x = 0$ and applied the continuity condition of the potential at $x = 0$. This potential is plotted in Fig. 4.7.

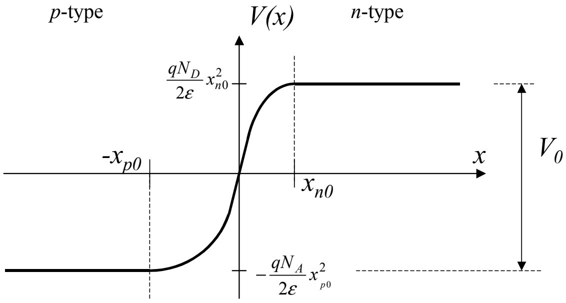

p-type
$V(x)$
n-type
$\frac{qN_D}{2\varepsilon}x_{n0}^2$
$-x_{p0}$
x
x_{n0}
V_0
$-\frac{qN_A}{2\varepsilon}x_{p0}^2$

Fig. 4.7. Built-in potential profile across a p-n junction. In the depletion approximation, there is no variation of the potential outside the depletion region.

The total potential difference across the *p-n* junction is called the built-in potential and is conventionally denoted V_{bi} or V_0. It can be obtained by evaluating the potential difference between $x = -x_{p0}$ and $x = x_{n0}$:

Eq. (4.17) $$V_0 = V(x_{n0}) - V(-x_{p0})$$

This can be rewritten as:

Eq. (4.18) $$V_0 = \frac{qN_D}{\varepsilon}\frac{x_{n0}^2}{2} + \frac{qN_A}{\varepsilon}\frac{x_{p0}^2}{2}$$

Expressing $-x_{p0}$ and x_{n0} as a function of the depletion width given in Eq. (4.8), we obtain:

Eq. (4.19) $$V_0 = \frac{q}{2\varepsilon}\frac{N_A N_D}{(N_A + N_D)}W_0^2$$

Another independent expression of the built-in potential can be obtained by expressing the balancing of the diffusion and drift currents. The total current from the motion of holes and that from the motion of electrons are given by:

Eq. (4.20) $$\begin{cases} J_h^{diff} + J_h^{drift} = -qD_p \dfrac{dp(x)}{dx} + q\mu_h p(x)E(x) \\ J_e^{diff} + J_e^{drift} = qD_n \dfrac{dn(x)}{dx} + q\mu_e n(x)E(x) \end{cases}$$

In these expressions, *p(x)* and *n(x)* represent the hole and electron concentrations at a position *x*. Taking into account the condition of Eq. (4.3) stating the exact balancing of the diffusion and drift currents for holes and electrons, we can write:

Eq. (4.21) $$\begin{cases} D_p \dfrac{dp(x)}{dx} = \mu_h p(x)E(x) \\ D_n \dfrac{dn(x)}{dx} = -\mu_e n(x)E(x) \end{cases}$$

which can be rewritten using Eq. (4.15) as:

Eq. (4.22) $$\begin{cases} \dfrac{D_p}{\mu_h}\dfrac{1}{p(x)}\dfrac{dp(x)}{dx} = -\dfrac{dV(x)}{dx} \\ \dfrac{D_n}{\mu_e}\dfrac{1}{n(x)}\dfrac{dn(x)}{dx} = \dfrac{dV(x)}{dx} \end{cases}$$

By integrating these equations, we get successively:

Eq. (4.23) $$\begin{cases} \dfrac{D_p}{\mu_h}\displaystyle\int_{-x_{p0}}^{x_{n0}} \dfrac{1}{p(x)}\dfrac{dp(x)}{dx}dx = -\int_{-x_{p0}}^{x_{n0}} \dfrac{dV(x)}{dx}dx \\ \dfrac{D_n}{\mu_e}\displaystyle\int_{-x_{p0}}^{x_{n0}} \dfrac{1}{n(x)}\dfrac{dn(x)}{dx}dx = \int_{-x_{p0}}^{x_{n0}} \dfrac{dV(x)}{dx}dx \end{cases}$$

Using Eq. (4.5) and Eq. (4.17), and by taking into account the Einstein relations: $\dfrac{D_p}{\mu_h} = \dfrac{D_n}{\mu_e} = \dfrac{k_bT}{q}$, we get:

Eq. (4.24) $$\begin{cases} \dfrac{k_bT}{q}\displaystyle\int_{p_p}^{p_n}\dfrac{dp}{p} = -\displaystyle\int_0^{V_0} dV \\ \dfrac{k_bT}{q}\displaystyle\int_{n_p}^{n_n}\dfrac{dn}{n} = \displaystyle\int_0^{V_0} dV \end{cases}$$

which integrates easily into:

Eq. (4.25) $$\begin{cases} \dfrac{k_bT}{q}\ln\left(\dfrac{p_n}{p_p}\right) = -V_0 \\ \dfrac{k_bT}{q}\ln\left(\dfrac{n_n}{n_p}\right) = V_0 \end{cases}$$

i.e.:

Eq. (4.26) $$V_0 = \frac{k_bT}{q}\ln\left(\frac{p_p}{p_n}\right) = \frac{k_bT}{q}\ln\left(\frac{n_n}{n_p}\right)$$

This can be rewritten into the form:

Eq. (4.27) $$\frac{p_p}{p_n} = \frac{n_n}{n_p} = \exp\left(\frac{qV_0}{k_bT}\right)$$

Using the expressions in Eq. (4.2), we can write the built-in potential as a function of the doping concentrations:

Eq. (4.28) $$V_0 = \frac{k_bT}{q}\ln\left(\frac{N_AN_D}{n_i^2}\right)$$

This potential exists at equilibrium and is a direct consequence of the junction between dissimilarly doped materials. However, it cannot be directly measured using a voltmeter because voltmeters measure the chemical potential difference, and the chemical potential is the same throughout the device since it is at thermal equilibrium with balanced drift and diffusion currents everywhere.

4.2.5. Depletion width

It is now possible to relate the width W_0 of the space charge region, as well as its extent on either side of the *p-n* junction, with the built-in potential. From the expression of the built-in potential in Eq. (4.19), we can express the depletion width as:

Eq. (4.29) $$W_0 = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)V_0}$$

which becomes, after considering Eq. (4.28):

Eq. (4.30) $$W_0 = \sqrt{\frac{2\varepsilon k_b T}{q^2}\left(\frac{N_A + N_D}{N_A N_D}\right)\ln\left(\frac{N_A N_D}{n_i^2}\right)}$$

The extent of the depletion width into each side of the *p-n* junction can then be determined by replacing W_0 from Eq. (4.29) into Eq. (4.8):

Eq. (4.31) $$\begin{cases} x_{p0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A + N_D)}\right)V_0} \\ x_{n0} = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A + N_D)}\right)V_0} \end{cases}$$

These last two expressions show that the space charge region extends more into the region of lower doping, in accordance with sub-section 4.2.2.

4.2.6. Energy band profile and Fermi energy

Because of the presence of a built-in potential, the allowed energy bands in the semiconductor, e.g. the conduction and the valence bands in particular, are shifted too. The resulting energy band profile is obtained by multiplying the potential by the charge of an electron ($-q$). This is shown in Fig. 4.8, where it is conventional to plot the bottom of the conduction band (E_C) and the top of the valence band (E_V) across the *p-n* structure.

The reason why we must multiply by the negative charge of an electron is because the resulting band diagram corresponds to the allowed energy states for electrons. This is intuitively understandable because the electrons are more likely to be where there is a higher positive electrical potential, thus the energy band for electrons will be lower there.

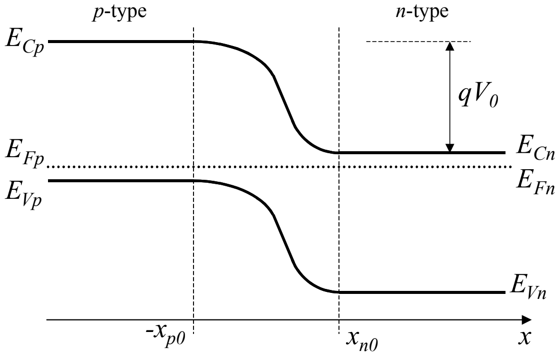


Fig. 4.8. Energy band profile across a p-n junction. This profile is obtained by multiplying the potential in Fig. 4.7 by −q, the electrical charge of electrons.

We therefore see that the conduction and valence bands are “bent” from the *p*-type to the *n*-type regions. Moreover, the amount of band bending is directly related to the built-in potential:

Eq. (4.32) $$E_{Vp} - E_{Vn} = E_{Cp} - E_{Cn} = qV_0$$

Away from the space charge region, the Fermi energy in the *p*-type and *n*-type regions are denoted E_{Fp} and E_{Fn} respectively, as shown in Fig. 4.8. At equilibrium, these quantities must be equal. Indeed, in the non-degenerate case the hole densities in the *p*-type and *n*-type regions are:

Eq. (4.33) $$\begin{cases} p_p = N_v \exp\left(\dfrac{E_{Vp} - E_{Fp}}{k_b T}\right) \\ p_n = N_v \exp\left(\dfrac{E_{Vn} - E_{Fn}}{k_b T}\right) \end{cases}$$

Utilizing Eq. (4.27), we get:

Eq. (4.34) $$\exp\left(\frac{qV_0}{k_b T}\right) = \frac{p_p}{p_n} = \frac{\exp\left(\dfrac{E_{Vp} - E_{Fp}}{k_b T}\right)}{\exp\left(\dfrac{E_{Vn} - E_{Fn}}{k_b T}\right)}$$

Eq. (4.35) $$\exp\left(\frac{qV_0}{k_b T}\right) = \exp\left(\frac{E_{Vp} - E_{Vn}}{k_b T}\right)\exp\left(\frac{E_{Fn} - E_{Fp}}{k_b T}\right)$$

In addition, by using Eq. (4.32) in this expression, we get:

Eq. (4.36) $$1 = \exp\left(\frac{E_{Fn} - E_{Fp}}{k_b T}\right)$$

which means that $E_{Fn} = E_{Fp}$, i.e. the Fermi energy in the *p*-type and *n*-type regions are equal and this has already been anticipated in Fig. 4.8. In fact, this is a general and important property that: at thermal equilibrium, the Fermi energies of dissimilar materials must be equal. This physically means that there must not be a net flow of holes or electrons across the structure at equilibrium.

4.3. Non-equilibrium properties of *p-n* junctions

The most interesting and practical properties of a *p-n* junction are observed under non-equilibrium conditions, such as when a voltage is applied across it (See Chapter 6) and/or when it is illuminated (see Chapter 9). Because of its non-symmetrical nature, a *p-n* junction will exhibit different properties depending on the polarity of the external voltage or bias applied. The sign convention used for the external voltage and the current in a *p-n* junction is shown in Fig. 4.9: the voltage will be positive if the applied potential on the *p*-type side is higher than that applied on the *n*-type. Note that the built-in voltage V_0 has been taken to be positive.

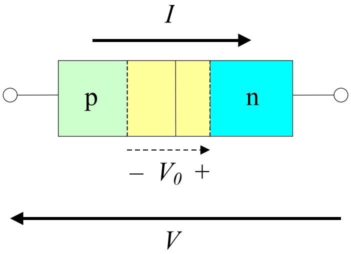


Fig. 4.9. Convention for the polarity of the external voltage and current.

When an external bias is applied, the diffusion and drift currents do not balance each other anymore. This imbalance results in a net flow of electrical current in one or the other direction. In addition, the internal electric field and voltage across the *p-n* junction, the depletion width and the energy band profile will all be changed. In this section, we will review how these parameters are modified.

4.3.1. Forward bias: a qualitative description

When an external bias V is applied to the *p-n* structure depicted in Fig. 4.9, there is usually some voltage drop across both the neutral bulk *p*-type and the *n*-type regions (i.e. outside the space charge region) due to Ohm's law. In other words, the entire external bias is not applied across the transition region because part of it would be "lost" across the neutral regions due to their electrical resistance.

However, in most semiconductor devices which use *p-n* junctions, the length of these neutral regions which the electrical current would have to flow through is small, and any voltage drop would thus be negligible compared to the voltage change across the transition region. In our discussion, for now we will assume that the external bias is applied directly to the limits of the space charge region.

According to the sign convention in Fig. 4.9, the total voltage across the transition region is now given by V_0-V. There are typically two regimes which need to be considered for the non-equilibrium conditions of a *p-n* junction: forward bias and reverse bias.

In the forward bias regime, corresponding to $V > 0$, the total voltage or potential barrier across the transition region is actually reduced from V_0 to V_0-V, which has a number of consequences. First, the strength of the internal electric field associated with the lower potential barrier is reduced as well, as shown in Fig. 4.10(c). This in turn means that the width of the space charge region is reduced because fewer electrical charges are needed to maintain this electric field, as shown in Fig. 4.10(b). In other words, W_0 is reduced and is now denoted W, x_{n0} becomes x_n, and x_{p0} becomes x_p, as illustrated in Fig. 4.10(a). As the internal voltage is reduced from its equilibrium value by an amount equal to V, the energy band profile is changed and the amount of band bending is reduced by qV, as depicted in Fig. 4.10(e). This means that:

Eq. (4.37) $$E_{Vp}-E_{Vn}=E_{Cp}-E_{Cn}=q(V_0-V)$$

instead of Eq. (4.32). Furthermore, we can still consider that the Fermi energy levels outside the space charge region, i.e. in the neutral bulk *p*-type (E_{Fp}) and *n*-type (E_{Fn}) regions, are located at their equilibrium positions because we assumed no voltage drop in these regions. Therefore, because the band bending has been reduced by qV, according to Fig. 4.10(e), we must have:

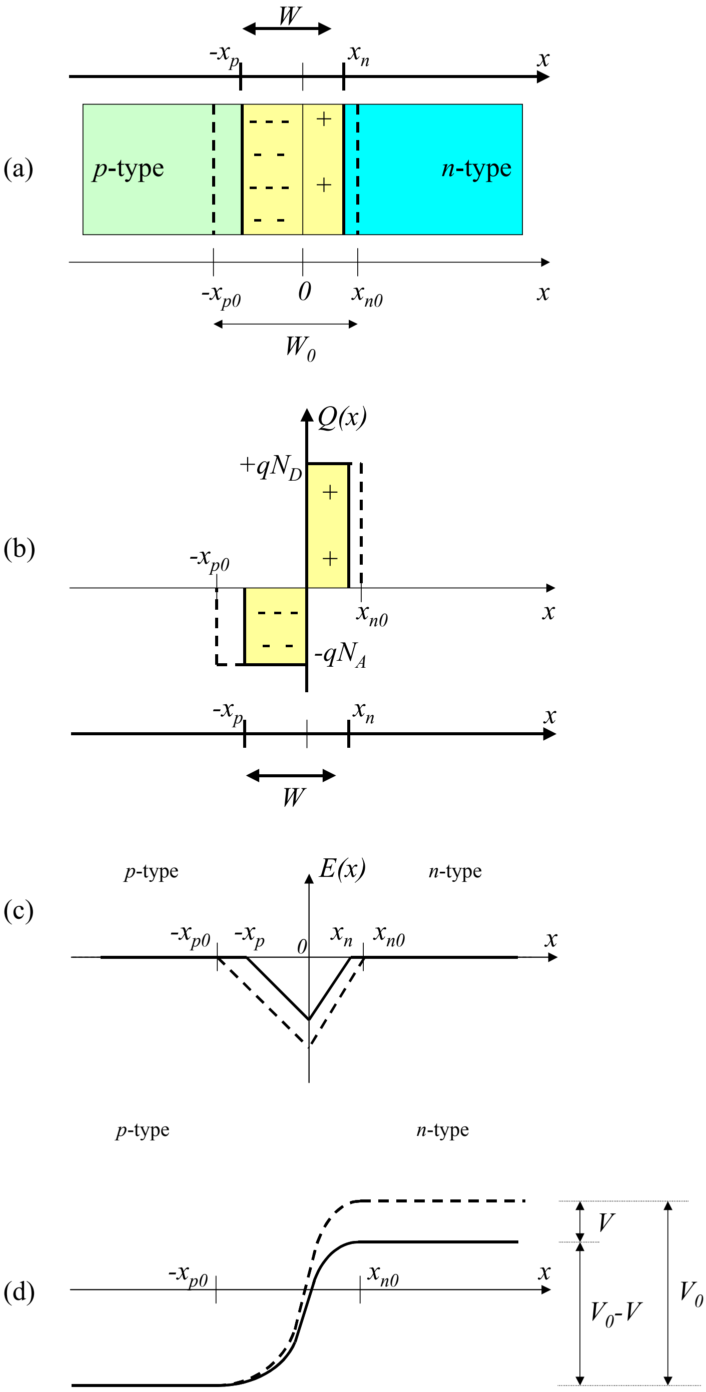
Eq. (4.38) $$E_{Fp}-E_{Fn}=-qV$$

This means that the Fermi energy is not constant throughout the *p-n* junction structure, but the Fermi energy levels in the neutral *p*-type and the *n*-type regions are separated by qV, where V is the applied external bias. This is a direct consequence of a non-equilibrium condition.

Let us now qualitatively examine the effects of a forward bias on the diffusion and drift currents across the space charge region of a *p-n* junction. As we saw in the previous section, the diffusion current arises from the difference between the density of charge carriers on either side of the junction area. It corresponds to the motion of electrons from the *n*-type region toward the *p*-type region, and conversely for holes. This means that, at its origin, the diffusion current is related to the motion of majority carriers (e.g. electrons in the *n*-type region). However, as soon as these carriers reach the other side of the junction, they become minority carriers. Therefore, the diffusion current acts as if it injects minority carriers into one side of the junction by pulling them from the other side of the junction where they are majority carriers.

At equilibrium, the diffusion process is stabilized when the built-in electric field exerts a force that exactly counterbalances the diffusion of these charge carriers. Under a forward bias, as we just saw in Fig. 4.10(c), this electric field strength is reduced. Therefore, each type of charge carriers can diffuse more easily, which means that the diffusion currents for both types of carrier increase under a forward bias.

This can also be understood by examining the energy band profile. For example: when the electrons in the *n*-type region, on the right hand side of Fig. 4.10(e) where they are more concentrated, diffuse towards the *p*-type region where they are less concentrated, the allowed energy states are located at higher energies. This means that the diffusion electrons have to cross a high-energy barrier. Under a forward bias, this energy barrier is reduced, as shown in Fig. 4.10(e), and more electrons can thus participate in the diffusion towards the *p*-type region. A similar argument is valid for holes. As a result, *the diffusion currents for both types of carrier increase under a forward bias*.


(a)
W
$-x_p$
x_n
x
p-type
n-type
$-x_{p0}$
0
x_{n0}
x
W_0
(b)
$Q(x)$
$+qN_D$
$-x_{p0}$
x_{n0}
x
$-qN_A$
$-x_p$
x_n
x
W
(c)
p-type
$E(x)$
n-type
$-x_{p0}$
$-x_p$
0
x_n
x_{n0}
x
p-type
n-type
(d)
$-x_{p0}$
x_{n0}
x
V
V_0-V
V_0

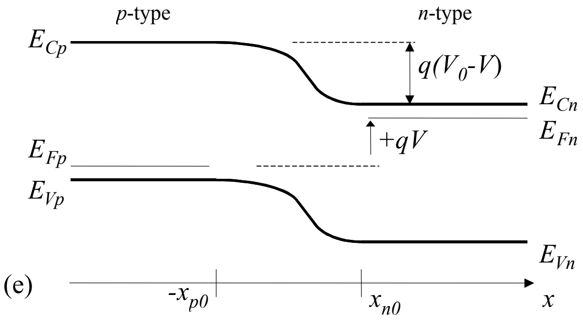
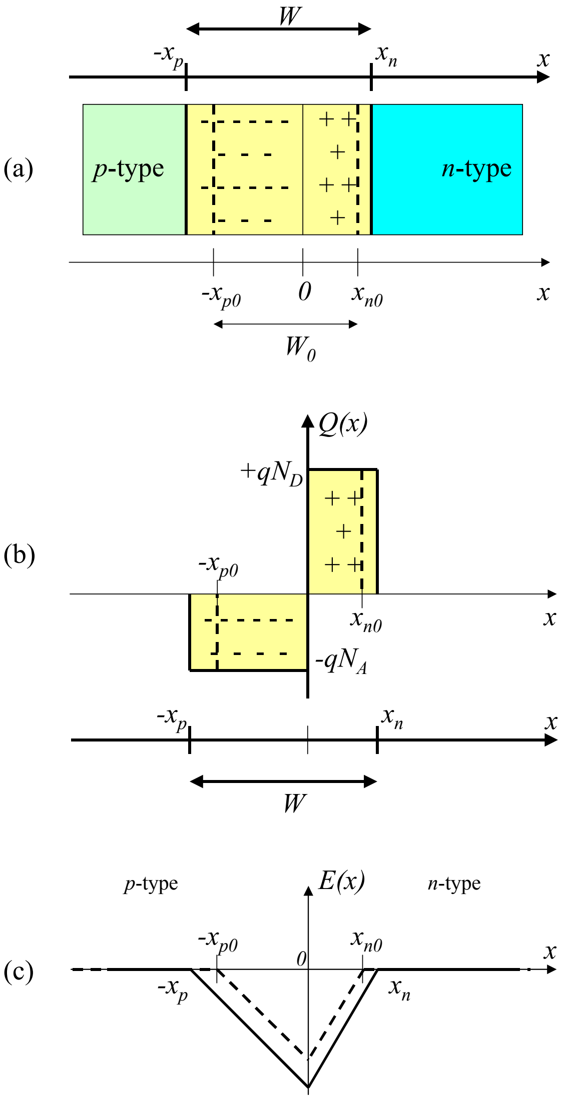


Fig. 4.10. (a) Space charge region width, (b) electrical charge density, (c) electric field strength, (d) potential profile, and (e) energy band profile of a p-n junction under forward bias (V>0). The thick dashed curves represent the equilibrium case for comparison.

By contrast, the *drift current does not change with an external bias*, although this may seem contradictory with the fact that the internal electric field is weaker. This can be understood by examining the drift current in more detail. We saw in Section 4.2 that the drift current counterbalanced the diffusion of charge carriers and thus consisted of electrons moving toward the *n*-type region and holes moving toward the *p*-type region. This means that, at its origin, the drift current is related to the motion of minority carriers, such as electrons in the *p*-type region which drift toward the *n*-type region under the influence of the electric field. The drift current thus plays the converse role of the diffusion current. The drift current acts as if it extracts minority carriers from one side of the junction to send them to the other side of the junction where they are majority carriers. Because the concentrations of minority carriers are very small (see Eq. (4.2)), the drift currents are mostly limited by the number of minority carriers available for drift (i.e. electrons on the *p*-type region and holes on the *n*-type region) rather than by the speed at which they would drift (i.e. the strength of the electric field). We then understand why the drift current does not change significantly when an external bias is applied, in comparison to the diffusion current.

4.3.2. Reverse bias: a qualitative description

By contrast, in the reverse bias regime, corresponding to $V < 0$, the total voltage or potential barrier across the transition region is actually increased from V_0 to $V_0 - V$, which also has the opposite effects of a forward bias. The strength of the internal electric field is increased, as shown in Fig. 4.11(c). This enlarges the width of the space charge region from W_0 to W (with x_{n0} becoming x_n, and x_{p0} becoming x_p, as illustrated in Fig. 4.11(a)) because more electrical charges are needed to maintain this electric field, as shown in Fig. 4.11(b). As the internal voltage is increased from its equilibrium value by an amount equal to $-V$, the energy band profile is changed and the amount of band bending is increased by $-qV$, as depicted in Fig. 4.11(e). The total amount of band bending is still given by the expression in Eq. (4.37). The difference between the Fermi energy levels outside the space charge region is also still given by Eq. (4.38).


W
$-x_p$
x_n
x
(a)
p-type
n-type
$-x_{p0}$
0
x_{n0}
x
W_0
Q(x)
$+qN_D$
(b)
$-x_{p0}$
x_{n0}
x
$-qN_A$
$-x_p$
x_n
x
W
p-type
E(x)
n-type
$-x_{p0}$
x_{n0}
0
x
(c)
$-x_p$
x_n

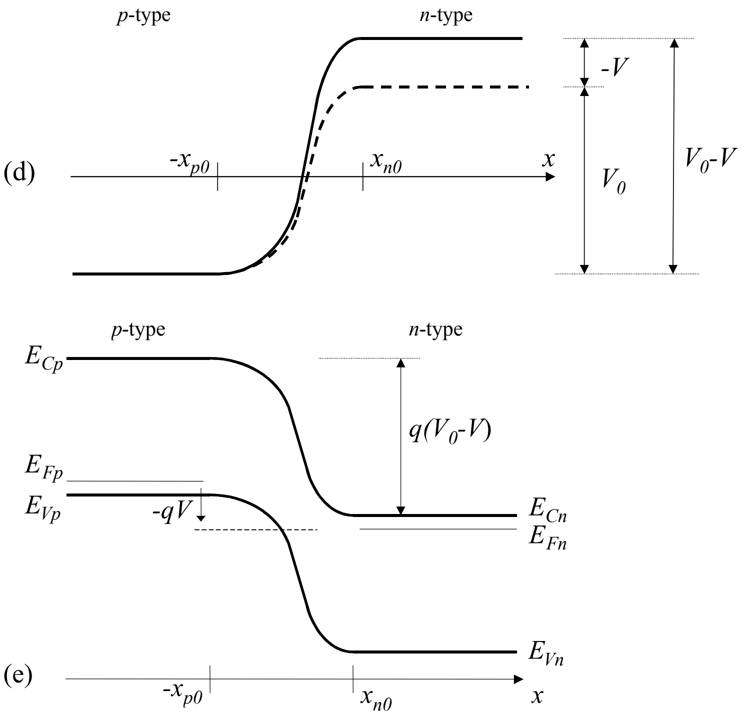


Fig. 4.11. (a) Space charge region width, (b) electrical charge density, (c) electric field strength, (d) potential profile, and (e) energy band profile of a p-n junction under reverse bias (V<0). The thick dashed curves represent the equilibrium case for comparison.

In addition, by contrast with the forward bias case, *the diffusion currents for both types of carrier decrease under a reverse bias*. However *the drift current still does not change significantly* in comparison to the diffusion current when a reverse bias is applied, for the same reason as discussed previously.

4.3.3. A quantitative description

In the previous sub-sections, we have expressed quantitatively the amount of band bending and the difference between the Fermi energy levels of the neutral *p*-type and *n*-type regions as a function of the applied external bias (Eq. (4.37) and Eq. (4.38) respectively).

In fact, most of the relations that were derived in section 4.2 for the equilibrium case are valid when an external bias voltage V is applied, provided we make the following transformations:

Eq. (4.39)
$$\begin{cases} W_0 & \rightarrow & W \\ x_{p0} & \rightarrow & x_p \\ x_{n0} & \rightarrow & x_n \\ V_0 & \rightarrow & V_0 - V \end{cases}$$

This statement is justified by the fact that most of the expressions in section 4.2 have been obtained without invoking the equilibrium condition of Eq. (4.3), but by using the electrical charge neutrality principle and Gauss's law instead which are valid at all times.

The following few relations will be important for future discussions. The depletion width can be obtained from Eq. (4.29) by using Eq. (4.39):

Eq. (4.40)
$$W = \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)(V_0 - V)}$$

for $V < V_0$. We clearly see that the depletion width shrinks when a forward bias is applied ($V > 0$) whereas it expands when a reverse bias is applied ($V < 0$). This confirms the qualitative discussion of the previous sub-section.

The extent of the space charge region inside the *p*-type and *n*-type regions, as shown in Fig. 4.10(a) and Fig. 4.11(a), can be obtained from Eq. (4.31):

Eq. (4.41)
$$\begin{cases} x_p = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_D}{N_A(N_A + N_D)}\right)(V_0 - V)} \\ x_n = \sqrt{\dfrac{2\varepsilon}{q}\left(\dfrac{N_A}{N_D(N_A + N_D)}\right)(V_0 - V)} \end{cases}$$

Similarly, the non-equilibrium hole and electron concentrations at the edges of the space charge region, denoted $p(-x_p)$, $p(x_n)$, $n(-x_p)$ and $n(x_n)$, can be obtained by considering Eq. (4.27):

Eq. (4.42)
$$\frac{p(-x_p)}{p(x_n)} = \frac{n(x_n)}{n(-x_p)} = \exp\left(\frac{q(V_0 - V)}{k_b T}\right)$$

In addition, following our previous discussion, we realize that the majority carrier concentrations is little changed under a moderate forward or a reverse bias, i.e. $p(-x_p) = p_p$ and $n(-x_n) = n_n$, which after replacing in Eq. (4.42) to:

Eq. (4.43) $$\frac{p_p}{p(x_n)} = \frac{n_n}{n(-x_p)} = \exp\left(\frac{q(V_0 - V)}{k_b T}\right)$$

and by using Eq. (4.27) to eliminate p_p and n_n from this latest equation:

Eq. (4.44) $$\frac{p(x_n)}{p_n} = \frac{n(-x_p)}{n_p} = \exp\left(\frac{qV}{k_b T}\right)$$

These expressions are important as they show that, when an external bias voltage is applied, the *minority carrier concentrations* at the boundary of the space charge region, $p(x_n)$ and $n(x_p)$, are directly and simply related to the equilibrium minority carrier concentrations p_n and n_p, and the applied bias voltage V. All these relations will prove important in the derivation of the diode equation for an ideal *p-n* junction which will be the topic of the next sub-section.

4.3.4. Ideal p-n junction diode equation

The diode equation refers to the mathematical expression which relates the total electrical current I through an ideal *p-n* junction to the applied external bias voltage V. It is also referred as the current-voltage or *I-V* characteristic of the diode. To determine it, we must focus our analysis on the minority carriers, i.e. holes in the *n*-type region and electrons in the *p*-type region.

In addition to the depletion approximation model considered so far, a few more assumptions need to be considered. (i) First, we assume that there are no external sources of carrier generation. (ii) No recombination of charge carriers occurs within the space charge region. (iii) We assume that the applied biases are moderate enough to ensure that the minority carriers remain much less numerous than the majority carriers in the neutral regions. (iv) Finally, we assume that the change in minority carrier concentrations in the neutral regions does not result in a non-negligible electric field.

In virtue of assumptions (i) and (ii), any hole or electron that has diffused across the space charge region must be present at its boundaries, i.e. at $-x_p$ and x_n respectively. When a bias V is applied, the concentrations of these holes and electrons, which are in excess of their equilibrium concentrations, are given by:

Eq. (4.45) $$\begin{cases} \Delta p_n = p(x_n) - p_n \\ \Delta n_p = n(-x_p) - n_p \end{cases}$$

This becomes after using Eq. (4.44):

Eq. (4.46)
$$\begin{cases} \Delta p_n = p_n\left(e^{\frac{qV}{k_bT}} - 1\right) \\ \Delta n_p = n_p\left(e^{\frac{qV}{k_bT}} - 1\right) \end{cases}$$

Here, and in the rest of the text, we will use the extended meaning of the term "excess carrier". For example, if Δp_n and Δn_p are positive, i.e. $V > 0$ or forward bias, then there are net real excesses of holes and electrons at the space charge boundaries and we talk about minority carrier injection. This is shown in Fig. 4.12

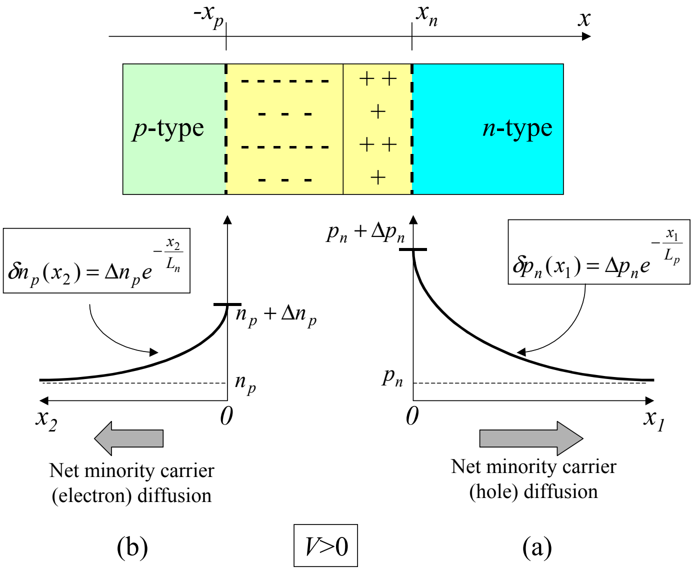

$-x_p$
x_n
x
p-type
n-type
$\delta n_p(x_2) = \Delta n_p e^{-\frac{x_2}{L_n}}$
$n_p + \Delta n_p$
n_p
x_2
0
$p_n + \Delta p_n$
$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}}$
p_n
0
x_1
Net minority carrier (electron) diffusion
Net minority carrier (hole) diffusion
(b)
$V>0$
(a)

Fig. 4.12. (a) Excess hole concentration profile in the n-type region, and (b) excess electron concentration profile in the p-type region, under a forward bias. The excess carrier concentrations decrease, following an exponential decay, as they go further from the edges of the depletion region.

But if Δp_n and Δn_p are negative, i.e. $V < 0$ or reverse bias, then there are net real deficiencies of holes and electrons and we talk about minority carrier extraction. In this case, the minority carriers at the boundaries of the space charge region are less numerous than in the bulk neutral material, therefore there is a diffusion of minority carriers from the bulk neutral region towards the edges of the space charge region. This is illustrated in Fig. 4.13.

Returning to the forward bias case, the excess holes, present at $x = x_n$ with a concentration Δp_n, will be diffusing deeper into the neutral *n*-type region where their equilibrium concentration is only p_n. As they diffuse, they will experience recombination with a characteristic diffusion length L_p in the steady-state regime. The excess hole concentration is therefore reduced as we advance deeper in the material. The analytical expression for $\delta p_n(x_1)$, the excess hole concentration at a position x_1, is:

Eq. (4.47) $$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}}$$

where L_p is the hole diffusion length in the *n*-type region. In this expression, we chose another axis, denoted x_1, oriented in the same direction as the original axis x and with its origin at $x = x_n$. It is important to remember that the excess concentration of holes at $x = x_n$ remains constant at Δp_n given by Eq. (4.46) because holes are continuously injected or extracted through the space charge region into or from the *n*-type region due to the application of the external bias voltage. We can plot the spatial profile of the excess hole concentration in Fig. 4.12(a) for the forward bias case and Fig. 4.13(a) for the reverse bias case.

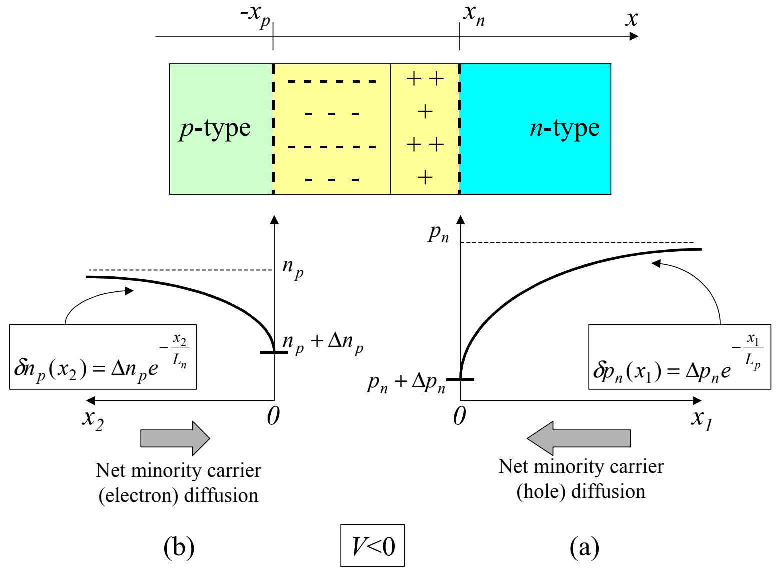

$-x_p$
x_n
x
p-type
n-type
n_p
$n_p + \Delta n_p$
$\delta n_p(x_2) = \Delta n_p e^{-\frac{x_2}{L_n}}$
x_2
0
p_n
$p_n + \Delta p_n$
$\delta p_n(x_1) = \Delta p_n e^{-\frac{x_1}{L_p}}$
0
x_1
Net minority carrier (electron) diffusion
Net minority carrier (hole) diffusion
(b)
$V<0$
(a)

Fig. 4.13. (a) "Excess" hole concentration profile in the n-type region, and (b) "excess" electron concentration profile in the p-type region, under a reverse bias. These carrier concentrations change following an exponential dependence as they go further away from the edges of depletion region.

Conversely, the excess electrons present at $x = -x_p$ with a concentration Δn_p will diffuse deeper into the neutral *p*-type region, with a diffusion length L_n. This leads to the spatial profile $\delta n_p(x_2)$ shown in Fig. 4.12(b) for the forward bias case and Fig. 4.13(b) for the reverse bias case, and it is analytically given by:

Eq. (4.48) $$\delta n_p(x_2) = \Delta n_p e^{-\frac{x_2}{L_n}}$$

where L_n is the electron diffusion length in the *p*-type region. It is important to note that, here, we chose the sign convention for the axis x_2 in the opposite direction of the original axis x because the electrons diffuse in this opposite direction.

There are essentially two methods to compute the diode equation. The first one consists of analyzing the diffusion currents in the *p-n* junction. From our discussion in sub-section 4.3.1 we understand that, when an external bias is applied, the drift currents across the space charge region do not vary whereas the diffusion currents change. The sum of the increments in the hole and the electron diffusion currents across the space charge region is thus a direct measure of the net electrical current through the *p-n* junction since no net current is originally present at equilibrium, because we have assumed there are no external sources of carrier generation and because the total electrical current is constant throughout a two-terminal device, such as the *p-n* junction earlier shown in Fig. 4.9.

The incremental diffusion currents are the diffusion currents which result from the excess carriers in the material. The diffusion current densities for electrons and holes are given by:

Eq. (4.49) $$\begin{cases} J_h^{diff}(x_1) = -qD_p \dfrac{d(\delta p_n(x_1))}{dx_1} \\ J_e^{diff}(x_2) = qD_n \dfrac{d(\delta n_p(x_2))}{dx_2} \end{cases}$$

Using the expressions of the excess carrier concentrations in Eq. (4.46) and Eq. (4.48), we get:

Eq. (4.50) $$\begin{cases} J_h^{diff}(x_1) = +q\dfrac{D_p}{L_p}\Delta p_n e^{-\frac{x_1}{L_p}} \\ J_e^{diff}(x_2) = -q\dfrac{D_n}{L_n}\Delta n_p e^{-\frac{x_2}{L_n}} \end{cases}$$

In order to obtain the total current through the *p-n* junction, we must evaluate the diffusion current densities for holes and electrons at the limits of the space charge region at $x = x_n$ and $x = -x_p$ respectively, or equivalently at $x_1 = x_2 = 0$:

Eq. (4.51) $$\begin{cases} J_h^{diff}(0) = +q\dfrac{D_p}{L_p}\Delta p_n \\ J_e^{diff}(0) = -q\dfrac{D_n}{L_n}\Delta n_p \end{cases}$$

In all these expressions of current densities, it is important to remember that the sign convention for the current density $J_h^{diff}(x_1)$ is the same as the axis x, whereas for $J_e^{diff}(x_2)$ it is opposite that of axis x. The total current density is the sum of the hole and electron diffusion currents, with however a sign difference:

Eq. (4.52) $$J_{total} = J_h^{diff}(0) - J_e^{diff}(0)$$

The minus sign for $J_e^{diff}(0)$ accounts for the sign convention chosen for axis x_2. Inserting Eq. (4.51) into this relation, we get:

Eq. (4.53) $$J_{total} = q\left(\frac{D_p}{L_p}\Delta p_n + \frac{D_n}{L_n}\Delta n_p\right)$$

and using Eq. (4.46), we finally obtain:

Eq. (4.54) $$J_{total} = q\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)\left(e^{\frac{qV}{k_bT}} - 1\right)$$

The total current is given by the total current density multiplied by the area of the *p-n* junction. If we assume a uniform area *A*, we get:

Eq. (4.55) $$I_{total} = AJ_{total} = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)\left(e^{\frac{qV}{k_bT}} - 1\right)$$

By introducing a new term I_0, this can be rewritten as:

Eq. (4.56) $$I_{total} = I_0\left(e^{\frac{qV}{k_bT}} - 1\right)$$

with:

Eq. (4.57) $$I_0 = qA\left(\frac{D_p}{L_p}p_n + \frac{D_n}{L_n}n_p\right)$$

Eq. (4.56) and Eq. (4.57) represent the diode equation for an ideal *p-n* junction. This function is plotted in Fig. 4.14.

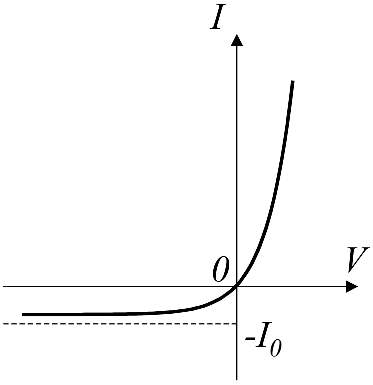


Fig. 4.14. Current-voltage characteristic for an ideal p-n junction diode. The dependence of the current on the voltage follows an exponential expression. The current is zero when the voltage is zero, without external excitation.

We see that under a forward bias, the current increases exponentially as a function of applied voltage. By contrast, under reverse bias, the current rapidly tends toward $-I_0$. The value of the current I_0 is therefore called the reverse saturation current. The physical meaning of this current can be understood as follows. When a strong reverse bias is applied (V<0), the density of minority carriers at the boundary of the space charge region quickly falls to zero according to Eq. (4.44). This means that, inside the depletion region, there is no diffusion of carriers but only drift currents are present. Outside the depletion region however, the only charge motion is the diffusion of minority carriers from the neutral regions toward the depletion region, as illustrated by the block arrows in Fig. 4.13. We can therefore say that the saturation current in Eq. (4.57) corresponds to the total drift, across the space charge region, of minority carriers which have been extracted or able to reach the limits of the space charge region through diffusion from the neutral regions.

The *p-n* junction diode acts like a one-way device: when it is forward biased, current can flow from the *p*-type to the *n*-type region without much resistance whereas when it is reverse-biased, a very large resistance prevents the current from flowing in the opposite direction from the *n*-type to the *p*-type region.

4.3.5. Minority and majority carrier currents in neutral regions

In the previous discussion, we saw that the total electrical current through a *p-n* junction device was determined by the diffusion currents across the space charge region which result in minority carriers being injected into or extracted from the neutral regions under the influence of an applied external bias.

For the sake of clarity, let us consider the example of a forward biased *p-n* junction, as the one shown in Fig. 4.12. We saw that the excess minority carriers diffuse into the neutral regions following an exponential decay given in Eq. (4.46) and Eq. (4.48). This leads to diffusion currents which also follow an exponential decay, as obtained in Eq. (4.50). However, we know that the total electrical current throughout a two-terminal device is constant. Therefore, the decrease in diffusion current, for example that of holes in the right hand side of the figure, as we move away from the space charge region has to be compensated by another current. This is achieved through the drift of majority carriers, for example electrons in the neutral *n*-type region. Indeed, through their diffusion and recombination, the minority carriers “consume” majority carriers (e.g. electrons). There thus must be a flow of majority carriers (e.g. electrons) in the opposite direction to resupply those lost in the recombination process. This flow of majority carriers generates a drift current.

Therefore, in the neutral regions, there are two components which make up the total electrical current: the diffusion current of minority carriers and the drift current of majority carriers. These are shown in Fig. 4.15. This means, in particular, that there must be an electric field present in the neutral regions, otherwise there would not be any drift current. This apparently contradicts our assumption at the beginning of sub-section 4.3.1 that there was no potential drop within the neutral regions. In fact, the potential drop is very small in comparison with any applied external bias voltage and therefore can be neglected in our model.

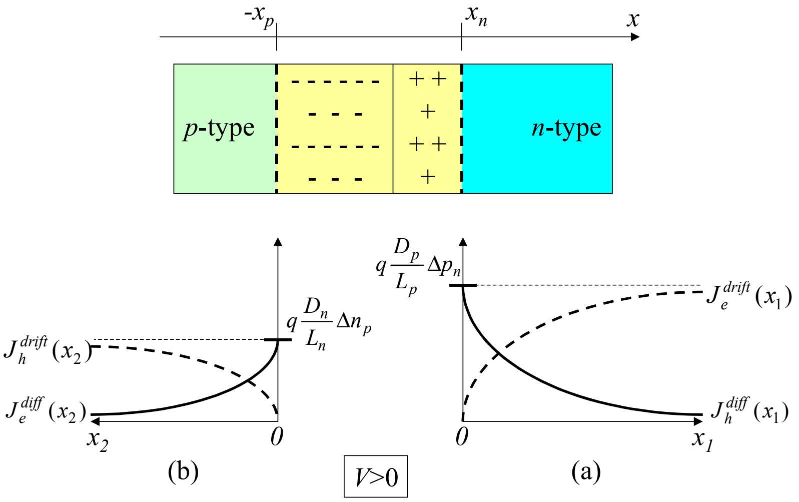


Fig. 4.15. Diffusion current of minority carriers and drift current of majority carriers in the (a) n-type region and (b) p-type region, under a forward bias. As the minority carriers diffuse further away from the edges of the depletion region, they recombine with majority carriers. The diffusion current of minority carriers is therefore reduced. But, this process also results in the flow of majority carriers in the opposite direction, which compensates the decrease in diffusion current with a drift current in the same proportion.

An analytical expression for the drift current can be easily determined, on each side of the *p-n* junction. Indeed, the total hole and electron current densities must be constant at the values given by the diode equation in Eq. (4.51). As we know the expression for the diffusion current densities $J_h^{diff}(x_1)$ and $J_e^{diff}(x_2)$ from Eq. (4.50), the drift current densities will be the difference between the total current and the diffusion currents on either side of the junction:

Eq. (4.58) $$\begin{cases} J_h^{drift}(x_2) = J_e^{diff}(0) - J_e^{diff}(x_2) \\ J_e^{drift}(x_1) = J_h^{diff}(0) - J_h^{diff}(x_1) \end{cases}$$

Recalling Eq. (4.50) and Eq. (4.53), we get successively:

Eq. (4.59)
$$\begin{cases} J_h^{drift}(x_2) = -q\dfrac{D_n}{L_n}\Delta n_p + q\dfrac{D_n}{L_n}\Delta n_p e^{-\frac{x_2}{L_n}} \\ J_e^{drift}(x_1) = q\dfrac{D_p}{L_p}\Delta p_n - q\dfrac{D_p}{L_p}\Delta p_n e^{-\frac{x_1}{L_p}} \end{cases}$$

Eq. (4.60)
$$\begin{cases} J_h^{drift}(x_2) = q\dfrac{D_n}{L_n}\Delta n_p\left(e^{-\frac{x_2}{L_n}} - 1\right) \\ J_e^{drift}(x_1) = q\dfrac{D_p}{L_p}\Delta p_n\left(1 - e^{-\frac{x_1}{L_p}}\right) \end{cases}$$

It is important to remember that the sign convention chosen for $J_h^{drift}(x_2)$ is opposite that of axis *x*.

4.4. Metal-semiconductor junctions

As we have already mentioned in sub-section 4.2.6 and illustrated in the case of a *p-n* junction, two dissimilar materials in contact with each other and under thermal equilibrium must have the same value of Fermi energy.

When a metal is brought into contact with a semiconductor, a certain amount of band bending occurs to compensate the difference between the Fermi energies of the metal and that of the semiconductor. In fact, this difference in Fermi energy means that electrons in one material have a higher energy than in the other. These will therefore tend to flow from the former to the later material. There is thus a transfer of electrons across the metal-semiconductor junction in a similar way as the charge transfer in the case of a *p-n* junction. Such a junction is also often called a metallurgic junction or a metal contact because metals are commonly used in semiconductor industry to connect or "contact" a semiconductor material to an external electrical circuit.

The charge transfer can be readily achieved because, the Fermi energy in a metal lies within an energy band, which makes it easy for electrons to be emitted from or received by a metal. This charge redistribution gives rise to a local built-in electric field which counterbalances this redistribution. When sufficiently large electric field strength is established around the metallurgic junction, the redistribution stops.

Since the overall charge neutrality must be maintained, the excess electrical charges inside the semiconductor and that inside the metal must be of an equal amount but with opposite signs. However, because a metal has a much higher charge density than a semiconductor, the width over which these excess charges spread inside the metal is negligibly thin in comparison to the width inside the semiconductor. This is somewhat similar to the case of a *p-n* junction with one side heavily doped. As a result, the built-in electric field and the band bending are primarily present inside the semiconductor as well. The following section aims at giving a quantitative description of the physical properties of a metal-semiconductor junction.

4.4.1. Formalism

The physical parameters which need to be considered in this description are depicted in Fig. 4.16. For the metal, these include its Fermi energy E_{Fm} and work function $\Phi_m>0$. The work function of a metal is the energy required to extract one electron from the metal surface and pull it into the vacuum. In a more quantitative manner, the work function is the energy difference between the Fermi energy and the vacuum level as shown in Fig. 4.16. For the semiconductor, the parameters of interest also include its Fermi energy E_{Fs}, its work function $\Phi_s>0$, and also its electron affinity $\chi>0$. The latter is the energy required to extract one electron from the conduction band of the semiconductor into the vacuum, and is given by the energy difference between the bottom of the conduction band and the vacuum level. A few values of electron affinity for elements in the periodic table are given in Fig. A.12 in Appendix A.3.

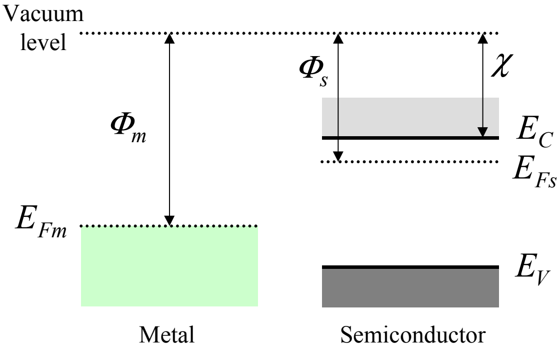


Fig. 4.16. Fermi energies, work functions in a metal and a semiconductor, when considered isolated from each other. The vacuum level is the same for both materials, but the Fermi energies are generally different.

The amount of band bending and the direction of electron transfer depend on the difference between the work functions of the metal and the semiconductor. When these materials are isolated, their vacuum levels are the same, as illustrated in Fig. 4.16. But, when these materials come into contact, the Fermi energy must be equal on both sides of the junction. The vacuum level is at an energy Φ_m above the top of the metal Fermi energy, while it is Φ_s above the semiconductor Fermi energy. This means that the energy bands in the semiconductor must shift upward by an amount equal to $\Phi_m - \Phi_s$ in order to align the Fermi energy on both sides of the junction.

On the one hand, if $\Phi_m > \Phi_s$, the energy bands of the semiconductor actually shift downward with respect to those of the metal and electrons are transferred from the semiconductor into the metal, as shown in Fig. 4.17. The signs of the charge carriers which appear on either side of the junction and the direction of the built-in electric field, also shown in Fig. 4.17, are determined from the analysis conducted for a *p-n* junction. On the other hand, if $\Phi_m < \Phi_s$, the energy bands in the semiconductor shift upward with respect to those of the metal and the electrons are transferred from the metal into the semiconductor.

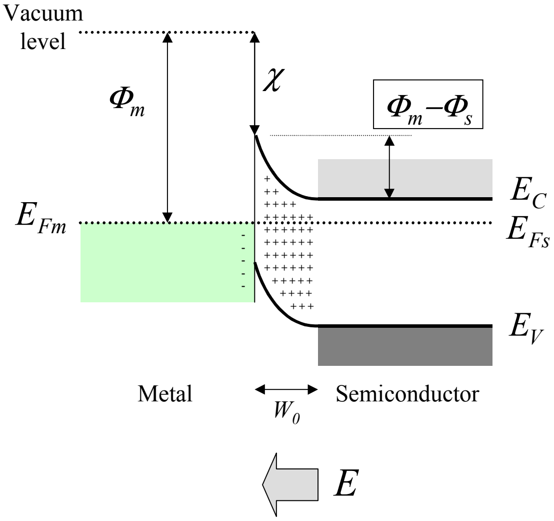


Fig. 4.17. Energy levels, accumulated charge carriers and built-in electric field in a metal-semiconductor junction. When the metal and the semiconductor are brought into contact, at equilibrium, the energy band profile of the semiconductor near the junction is modified so that the Fermi energies become equal in both materials.

4.4.2. Schottky and ohmic contacts

The electrical properties of a metal-semiconductor junction depend on whether a depletion region is created as a result of the charge redistribution. This phenomenon in turn depends on the difference in work function $\Phi_m - \Phi_s$, and on the type of the semiconductor (*n*-type or *p*-type).

Indeed, we know that when $\Phi_m > \Phi_s$, electrons are extracted from the semiconductor into the metal.

If the semiconductor is *n*-type, then this process depletes the semiconductor of its electrons or majority charge carriers. A depletion region thus appears near the junction and we obtain a diode-like behavior similar to a *p-n* junction when an external bias is applied. This is shown in Fig. 4.18(a). This situation is often called a rectifying contact or Schottky contact.

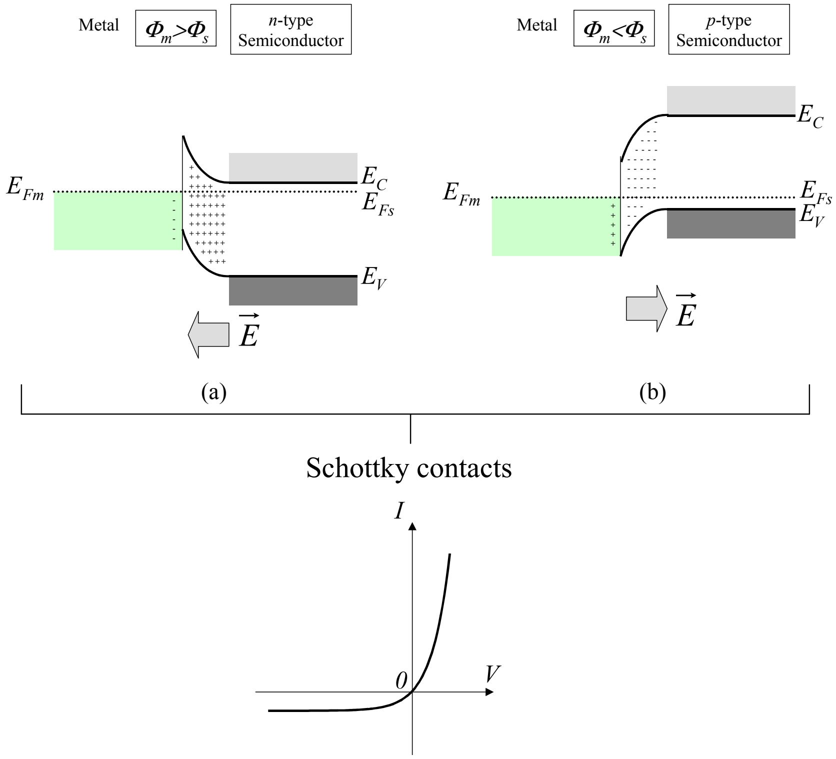


Fig. 4.18. These two of the four possible metal-semiconductor junction configurations lead to a Schottky contact: (a) $\Phi_m>\Phi_s$ and n-type, (b) $\Phi_m<\Phi_s$ and p-type. A Schottky contact is obtained in each case because the majority carriers in the semiconductor experience a potential barrier which prevents their free movement across the metal-semiconductor junction and therefore as shown at the bottom of the figure, the I-V characteristic shows rectifying behavior.

However, if the semiconductor is *p*-type, the electrons which are extracted from the semiconductor are taken from the *p*-type dopants which then become ionized. This process thus creates more holes or majority charge carriers. In this case, there is no depletion region, but rather majority carriers are accumulated near the junction area and we do not observe a diode-like behavior. Majority carriers are free to flow in either direction under the influence of an external bias. This is shown in Fig. 4.19(a). This situation is often called an ohmic contact and the current-voltage characteristics are linear.

If we now consider $\Phi_m < \Phi_s$, electrons are extracted from the metal into the semiconductor. The previous analysis needs to be reversed. In other words, for an *n*-type semiconductor the junction will be an ohmic contact, while for a *p*-type semiconductor the junction will be a Schottky contact.

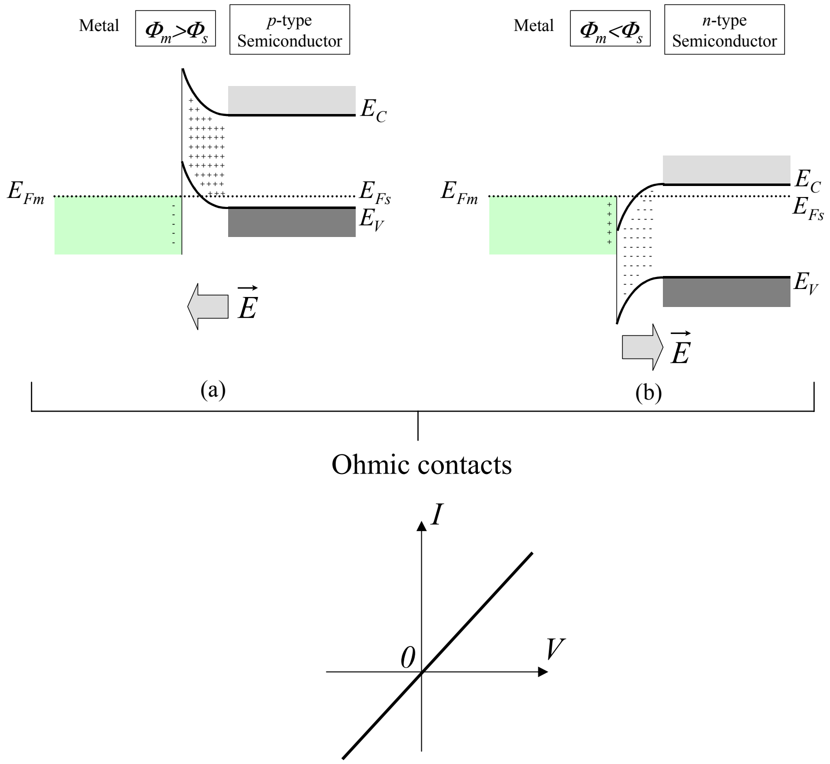


Fig. 4.19. These two of the four possible metal-semiconductor junction configurations lead to an ohmic contact: (a) $\Phi_m > \Phi_s$ and p-type, (b) $\Phi_m < \Phi_s$ and n-type. Unlike the configurations shown in Fig. 4.18, the energy band profiles here are such that the majority carriers in the semiconductor can move across the metal-semiconductor junction without experiencing a potential barrier and therefore as shown at the bottom of the figure, the I-V characteristic shows ohmic behavior.

These four configurations are shown in Fig. 4.18 and Fig. 4.19, and summarized in Table 4.1.

	Semiconductor	Junction
$\Phi_m > \Phi_s$	*n*-type	Schottky
$\Phi_m < \Phi_s$	*p*-type	Schottky
$\Phi_m > \Phi_s$	*p*-type	ohmic
$\Phi_m < \Phi_s$	*n*-type	ohmic

Table 4.1. Four possible metal-semiconductor junction configurations and the resulting contact types.

In the case of a Schottky contact, the existence of the depletion region means that there is a potential barrier across the junction which can be shifted by an amount equal to $-qV$ when an external voltage V is applied between the metal and the semiconductor. This in turn influences the current flow in a similar way as for a *p-n* junction. This is shown in Fig. 4.20 for the case of an *n*-type semiconductor. It is however important to understand that majority carriers are responsible for the current transport in a metal-semiconductor junction, whereas in a *p-n* junction it is due to the minority carriers.

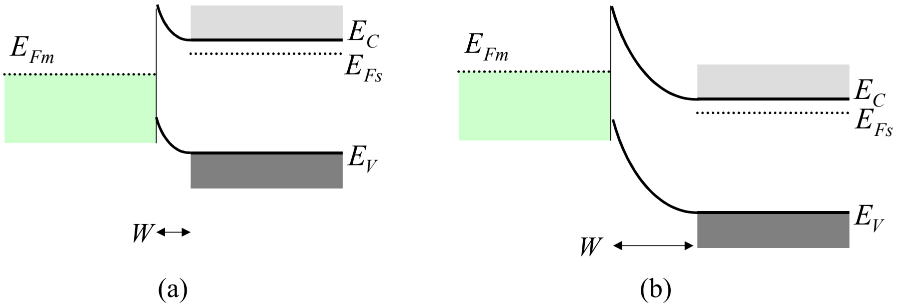


Fig. 4.20. Band alignment in a Schottky metal-(n-type) semiconductor contact under (a) forward bias where the potential barrier is reduced, and under (b) reverse bias where the potential barrier is increased, thus reducing the tunneling of carriers.

The sign convention for a metal-semiconductor junction is the same as for a *p-n* junction by considering the type of the semiconductor. Although the current transport mechanism in a Schottky contact is somewhat different from that in a *p-n* junction, the current-voltage relation for an ideal Schottky contact has a similar expression as for an ideal *p-n* junction:

Eq. (4.61)
$$I = I_0\left(e^{\frac{qV}{k_bT}} - 1\right)$$

where I_0 is the reverse saturation current and is exponentially proportional to the difference between the metal work function Φ_m and the semiconductor electron affinity χ:

Eq. (4.62)
$$I_0 = AB_eT^2e^{\left(-\frac{(\Phi_m-\chi)}{k_bT}\right)}$$

B_e is the effective Richardson constant, and for most metal-semiconductor Schottky junctions it varies from 10 to 100 K^{-2} cm^{-2} The quantity $(\Phi_m-\chi)$ is often denoted $q\Phi_B$, where Φ_B is called the Schottky potential barrier height. For a real Schottky contact, one needs to take into account thermionic emission (Appendix A.5), as well as impurity and interface states. In this case, the current-voltage relation is given by:

Eq. (4.63)
$$I = I_0\left(e^{\frac{qV}{nk_bT}} - 1\right)$$

where n is the ideality factor as mentioned before and is typically between 1 and 2.

4.5. Summary

In this Chapter, we have presented a complete mathematical model for an ideal *p-n* junction, based on an abrupt homojunction model and the depletion approximation. We introduced the concepts of a space charge region, built-in electric field, built-potential, and depletion width at equilibrium. We have discussed the balance of electrical charges, as well as that of the diffusion and drift currents within the space charge region.

The non-equilibrium properties of *p-n* junctions have also been discussed. The forward bias and reverse bias conditions were examined. We emphasized the importance of minority carrier injection and extraction. We derived the diode equation and understood the nature of the currents outside the space charge region.

Finally, we presented the electrical properties of metal-semiconductor junctions and introduced the concepts of Schottky and ohmic contacts.

Further reading

Ashcroft, N.W. and Mermin, N.D., *Solid State Physics*, Holt, Rinehart and Winston, New York, 1976.
Neudeck, G.W., *The PN Junction Diode*, Addison-Wesley, Reading, MA, 1989.
Pierret, R.F., *Advanced Semiconductor Fundamentals*, Addison-Wesley, Reading, MA, 1989.
Sapoval, B. and Hermann, C., *Physics of Semiconductors*, Springer-Verlag, New York, 1995.
Streetman, B.G., *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.
Wang, S., *Fundamentals of Semiconductor Theory and Device Physics*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Problems

1. A *p-n* junction diode has a concentration of $N_A = 10^{17}$ acceptor atoms per cm^3 on the *p*-type side and a concentration of N_D donor atoms per cm^3 on the *n*-type side. Determine the built-in potential V_0 at room temperature for a germanium diode for values of N_D ranging from 10^{14} to 10^{19} cm^{-3}. Also determine the peak value of the electric field strength for this same range, and plot both of these values as a function of N_D on a semilog scale.

2. Consider a GaAs step junction with $N_A = 10^{17}$ cm^{-3} and $N_D = 5\times10^{15}$ cm^{-3}. Calculate the Fermi energy in the *p*-type and *n*-type regions at 300 K. Draw the energy band diagram for this junction. Determine the built-in potential from the diagram and from Eq. (4.28). Compare the results.

3. Consider an asymmetric p^+-*n* junction, which has a heavily doped *p*-type side relative to the *n*-type side, i.e. $N_A \gg N_D$. Determine a simplified expression for the width of the space charge region given in Eq. (4.29).
   Calculate the depletion width for a Si *p-n* junction that has been doped with 10^{18} acceptor atoms per cm^3 on the *p*-type side and 10^{16} donor atoms per cm^3 on the *n*-type side. Compare this depletion

width to the width of the depletion region on the *n*-side (from Eq. (4.41)). What percentage of the width lies within the *n*-type semiconductor.

4. A silicon *p-n* diode with $N_A = 10^{18}\ \text{cm}^{-3}$ has a built-in voltage of 0.814 eV and capacitance of $10^{-8}\ \text{F}\cdot\text{cm}^{-2}$ at an applied voltage of 0.5 V. Determine the donor density.

5. Plot the diode equation for an ideal Si *p-n* junction diode with an area 50 μm^2, an acceptor concentration $N_A = 10^{18}\ \text{cm}^{-3}$, a donor concentration $N_D = 10^{18}\ \text{cm}^{-3}$, recombination lifetimes equal to $\tau_n = \tau_p = 1\ \mu\text{s}$, and diffusion coefficients equal to $D_n = 35\ \text{cm}^2\cdot\text{s}^{-1}$ and $D_p = 12.5\ \text{cm}^2\cdot\text{s}^{-1}$.

8. Consider an ideal metal-semiconductor junction between *p*-type silicon and polycrystalline aluminum. The Si is doped with $N_A = 5\times10^{16}\ \text{cm}^{-3}$. The metal work function is 4.28 eV and the Si electron affinity is 4.01 eV. Draw the equilibrium band diagram and determine the barriers height ϕ_B.

5. Transistors

5.1. Introduction
5.2. Overview of amplification and switching
5.3. Bipolar junction transistors
  5.3.1. Principles of operation for bipolar junction transistors
  5.3.2. Amplification process using BJTs
  5.3.3. Electrical charge distribution and transport in BJTs
  5.3.4. Current gain
  5.3.5. Typical BJT configurations
  5.3.6. Deviations from the ideal BJT case
5.4. Heterojunction bipolar transistors
  5.4.1. AlGaAs/GaAs HBT
  5.4.2. GaInP/GaAs HBT
5.5. Field effect transistors
  5.5.1. JFETs
  5.5.2. JFETs gate control
  5.5.3. JFET characteristics
  5.5.4. MOSFETs
  5.5.5. Deviations from the ideal MOSFET case
5.6. Application specific transistors
5.7. Summary

5.1. Introduction

Modern semiconductor electronics was revolutionized by the invention of the bipolar transistor at the Bell Telephone Laboratories in 1948. The impact of transistors can be best understood when one realizes that, without them, there would have been no progress in such diverse areas of everyday life as computers, television, telecommunications, the Internet, air travel, space exploration, as well as the tools necessary to study and understand the biological process.

Transistors can be classified in two major categories: they can be either bipolar transistors or field effect transistors. Each is fundamentally different

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_5,
© Springer Science+Business Media, LLC 2010

from the other in its operation mechanisms. A bipolar transistor operates through the injection and collection of *minority* carriers utilizing *p-n* junctions. By contrast, a field effect transistor is a *majority* carrier device and is thus a unipolar device.

In this Chapter, we will first review the general motivation and principles for electrical amplification and switching. We will then describe qualitative and quantitatively the direct-current (DC) operation mechanisms of bipolar junction transistors (BJT), their modes of operation, second order effects, as well as practical applications of BJTs in amplifier configurations. A variation of the bipolar transistor, utilizing a heterojunction, will subsequently be discussed. Next, the DC operating principles of field effect transistors (FET) will be presented along with corresponding second order behaviors of FET-based devices. Once again, practical circuit configurations and applications will be introduced. Finally, application specific transistors will be presented, including single electron as well as high electron mobility transistors.

5.2. Overview of amplification and switching

Transistors are capable of serving as switches and amplifiers, depending upon their configuration. The term “transistor” comes from “transfer resistor” and alludes to a transistor’s behavior as a resistor that amplifies signals as they are transferred from the input to the output terminal of the device.

Before trying to understand the amplification and switching mechanisms in a transistor, it is important to comprehend the idea of operating current and voltage of a given device. Let us consider the simple electrical circuit in Fig. 5.1(a), which includes a voltage source (e.g. a battery) V_0, a resistance R and the electrical device under consideration.

The current-voltage characteristic of the device under consideration is shown as the solid line in Fig. 5.1(b), which is the illustration of the mathematical function:

Eq. (5.1) $I_T=f(V_T)$

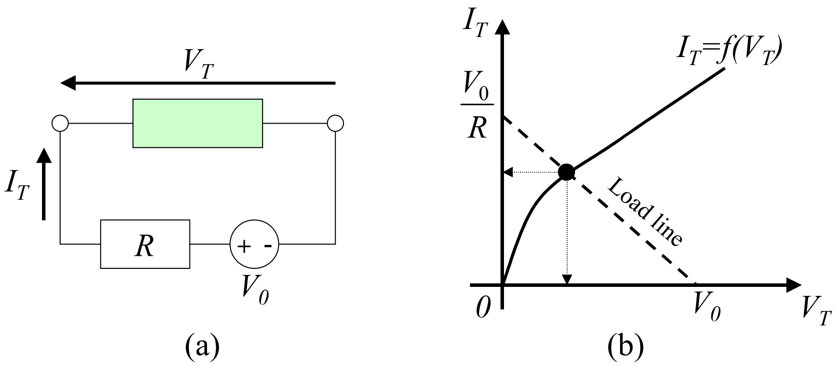


Fig. 5.1. (a) Example of electrical circuit. (b) Illustration of the current-voltage characteristic of the device (solid line) and the load line (dashed line).

In addition, the voltage-loop equation around the circuit shown in Fig. 5.1(a) yields:

Eq. (5.2) $$V_0 - RI_T = V_T$$

This equation is shown as the dashed line in Fig. 5.1(b). The steady state current I_T and voltage V_T are determined by solving the system formed by these two equations. The solution can be easily visualized graphically and corresponds to the intersection point of the two curves in Fig. 5.1(b).

Let us now consider an electrical device with three terminals or electrical connections, as shown in Fig. 5.2(a). Let us further assume that the current I_T through two of the terminals can be controlled by changes in the current $I_{control}$ or the voltage $V_{control}$ applied at the third terminal as illustrated in Fig. 5.2(b) by the collection of current-voltage characteristic. Eq. (5.2) is still valid and is still represented by the dashed line along which the steady state values of the current I_T and voltage V_T are located. This line is called the circuit load line.

As we can see graphically in Fig. 5.2(b), the significant changes in the current I_T (e.g. ~50 mA) can be achieved by only small changes in the control current $I_{control}$ (e.g. ~0.3 mA). This feature is called amplification, through which a small signal variation, such as that of $I_{control}$, can be amplified into a large signal change such as that of I_T.

Another important feature of this type of electrical circuit is the possibility to turn *on* and *off* the device through changes in $I_{control}$. This is achieved by switching the current I_T between the two extremes on the load lines, from $I_T = 0$ to $I_T = V_0/R$. This feature is called switching.

A transistor is an example of a three terminal device that exhibits amplification and switching capabilities. These are the basis of all electronic device functions, which makes transistors the basic elements in modern electronics.

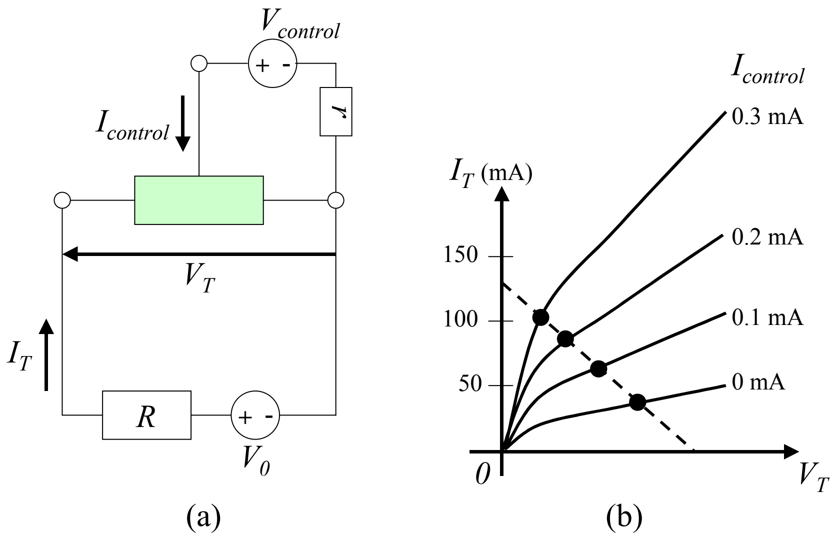


Fig. 5.2. (a) Example of electrical circuit utilizing a three terminal device. (b) Illustration of the load line (dashed line) and the current-voltage characteristic of the device (solid lines) as a function of the control current. The intersection of the solid and dashed lines gives the steady state values of current and voltage across the device.

5.3. Bipolar junction transistors

Bipolar junction transistors consist of two back-to-back *p-n* junctions which share a common terminal. Such a transistor can be a *p-n-p* or an *n-p-n* transistor. In this section, we shall use the *p-n-p* configuration for most illustrations and analysis. The main advantage of the *p-n-p* for discussing transistor action is that hole flow and current are in the same direction. This makes the various mechanisms of charge transport somewhat easier to visualize in a preliminary explanation. Once these basic ideas are established for the *p-n-p* device, it is simple to relate them to the more widely used *n-p-n* transistors. The corresponding current and bias polarities need only to be switched for the *n-p-n* case. The common schematic diagrams for *n-p-n* and *p-n-p* bipolar junction transistors are shown in Fig. 5.3. The direction of the arrow on the emitter leg indicates the direction of current flow, during the forward active mode of operation, and the *p-n* transition; thus, it can be used to easily identify the transistor type in circuit diagrams. We will start by discussing the BJT from a qualitative viewpoint and gain physical understanding on how the device operates.

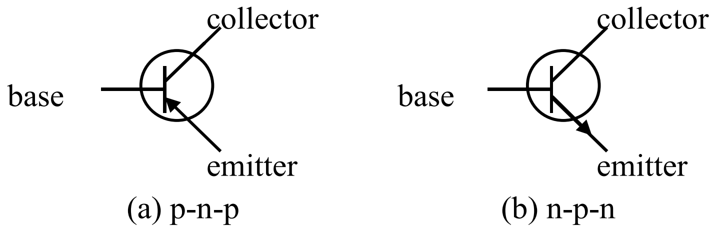


Fig. 5.3. Schematic diagrams for (a) p-n-p and (b) n-p-n bipolar junction transistors.

5.3.1. Principles of operation for bipolar junction transistors

We will now qualitatively describe how the current control can be achieved using a BJT. Let us consider the *p-n-p* transistor shown in Fig. 5.4 with the left *p-n* junction under forward bias and the right one under reverse bias, otherwise known as the forward active mode of operation.

According to the analysis of carrier transport in a *p-n* junction done in Chapter 4 (section 4.3.1), the left *p-n* junction is biased such that holes are injected from its heavily doped *p*-type region into its *n*-type region (current i_E) where they become minority carriers and will diffuse to reach the other side of the *n*-type region. The left *p*-type region electrode is therefore called the emitter. The width, W, of the base region is typically thinner than the minority carrier diffusion length, to maximize the number of minority carriers that diffuse across it and to minimize base recombination.

As the *n-p* junction on the right is under reverse bias, the electrical current flowing through it (I_C) is mostly determined by the drift current across the depletion region and its magnitude is determined by the concentration of minority carriers present at the boundaries, $x = W$ and $x = x_c$, of this depletion region.

Therefore, the holes which were injected by the *p-n* junction (left) and which succeeded to reach the edge of the base depletion, $x = W$, region for the *n-p* junction (right) via diffusion will determine the magnitude of the electrical current through the *n-p* junction. These holes are, in a way, collected by the right *p*-type region electrode, which is therefore called the collector. The I_C current corresponds to the saturation current given in Eq. (4.57) and is independent of the applied reverse bias voltage, neglecting any leakage.

The exact amount of holes that diffuse through the *n*-type region is affected by a few parameters which are intrinsic to the *n*-type region (e.g. diffusion lengths), as well as by other parameters which are extrinsic to it such as the current flowing through the base (I_B) which acts as the control current of Fig. 5.2. The mechanisms of this current control will now be illustrated in further details in the case of amplification in a bipolar junction transistor.

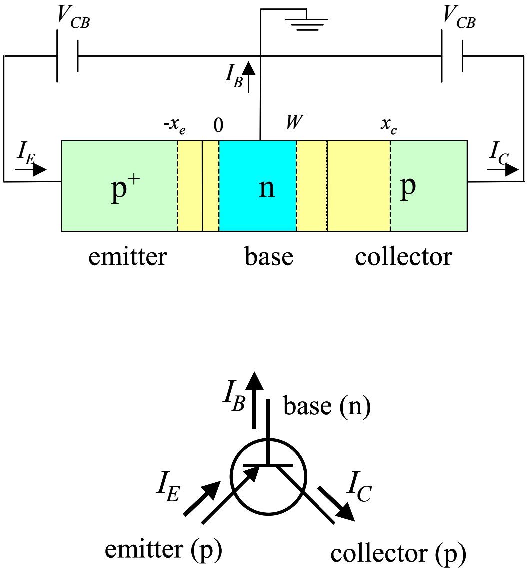


Fig. 5.4. A p-n-p bipolar junction transistor with its emitter-base junction under forward bias and the base-collector under reverse bias. The conventions for the signs of the electrical currents I_E, I_B and I_C are shown: the electrical current will be positive if it actually flows in the direction of the arrow.

5.3.2. Amplification process using BJTs

In order to understand how the amplification process is carried out, we will first develop a qualitative picture of the current mechanisms in a BJT transistor. Then, we will take a more analytical approach to expressing the current mechanisms, as well as the amplification and transport factors relevant to BJT devices.

Let us begin by considering a *p-n-p* transistor with a heavily doped p^+ region for the emitter. The *p*-type emitter is taken to be highly doped in order to be a more efficient hole emitter. A schematic diagram of such a structure is shown in Fig. 5.5.

For the holes that are injected from the emitter electrode into the base, a portion will undergo recombination with the electrons present in the base region (denoted as 1 in Fig. 5.5). The probability of recombination is proportional to the density of electrons available and the density of holes that are injected.

When no external electrons are injected into the base region, the recombining holes will lead to the apparition of fixed positive ions. The density of electrons will then decrease, thus causing less and less electron-hole recombination, and the fixed positive ions will build up. The resulting electric field will reduce the injection of the holes into the base region, and hence, the portion of the holes injected from the emitter and reaching the collector will decrease as well.

However, when electrons are injected into the base region through the base current I_B (4 in Fig. 5.5), they will recombine with the holes and reduce the buildup of positive charges. The base barrier will therefore decrease and a larger amount of holes will reach the collector (2 in Fig. 5.5).

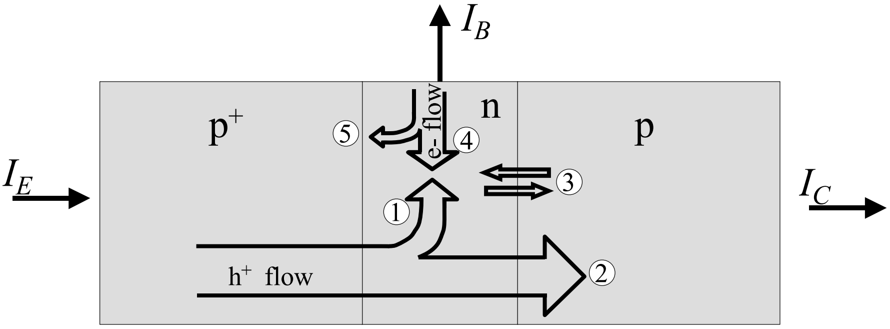


Fig. 5.5. Schematic diagram showing the flows of holes and electrons within a p^+-n-p bipolar junction transistor. The conventions for the signs of the electrical currents i_E, i_B and i_C are shown: the electrical current will be positive if it actually flows in the direction of the arrow.

This phenomenon provides us with a method to control the current flow from the emitter to the collector via the amount of electrons injected into the base. Since only a small portion of holes will be recombined with the injected electrons, we can use a small injection current (I_B) to control a much bigger current (I_C). And the current gain can be very high if the recombination rate is low, which can be done by engineering the base region adequately.

5.3.3. Electrical charge distribution and transport in BJTs

We will now try to quantitatively examine the operation of bipolar junction transistors. We will consider a *p-n-p* transistor shown Fig. 5.6. The case of forward active operation will be considered. The objective will be to determine the minority carrier distributions and the terminal currents. In order to simplify the calculations, a few assumptions are made:

(1) Holes diffuse from the emitter to the collector and drift is negligible in the base region.
(2) The emitter current is contributed entirely by holes, i.e. the emitter injection efficiency γ is 1 and $I_{En} = 0$.
(3) The collector saturation current is negligible, i.e. component 3 in Fig. 5.5.
(4) The active part of the base and the two junctions are of uniform cross-sectional area A and the current flow in the base is essentially one-dimensional from the emitter to the collector.
(5) All currents and voltages are considered at steady state.

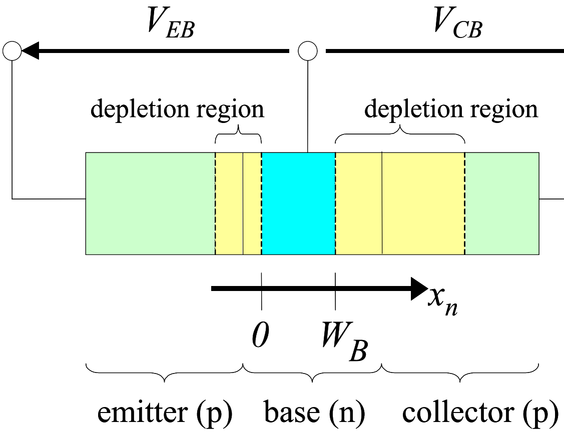


Fig. 5.6. Schematic diagram of a p-n-p bipolar junction transistor showing the voltages and convention for the position variable x_n.

The excess hole concentration on the collector side of the base Δp_C and that on the emitter side of the base Δp_E are given by:

$$\text{Eq. (5.3)} \quad \begin{cases} \Delta p_E = p_n(e^{\frac{qV_{EB}}{k_bT}} - 1) \\ \Delta p_C = p_n(e^{\frac{qV_{CB}}{k_bT}} - 1) \end{cases}$$

where p_n is the equilibrium hole concentration in the n-type base region. If the emitter junction is strongly forward biased ($V_{EB} >> k_bT/q$) and the collector junction is strongly reverse biased ($V_{CB} << 0$), these expressions can be simplified and become:

Eq. (5.4)
$$\begin{cases} \Delta p_E \approx p_n e^{\frac{qV_{EB}}{k_b T}} \\ \Delta p_C \approx -p_n \end{cases}$$

The diffusion equation is given by:

Eq. (5.5)
$$\frac{d^2 \delta p(x_n)}{dx_n^2} = \frac{\delta p(x_n)}{L_p^2}$$

where $\delta p(x_n)$ is the concentration of excess holes at x_n, and L_p is the hole diffusion length in the *n*-type base region. The general solution of this equation is:

Eq. (5.6)
$$\delta p(x_n) = C_1 e^{\frac{x_n}{L_p}} + C_2 e^{-\frac{x_n}{L_p}}$$

where C_1 and C_2 are integration constants. Expressing the boundary conditions, we get:

Eq. (5.7)
$$\begin{cases} \delta p(x_n = 0) = C_1 + C_2 = \Delta p_E \\ \delta p(x_n = W_b) = C_1 e^{\frac{W_b}{L_p}} + C_2 e^{-\frac{W_b}{L_p}} = \Delta p_C \end{cases}$$

where W_b is the width of the base region. Solving for C_1 and C_2 we get:

Eq. (5.8)
$$\begin{cases} C_1 = \dfrac{\Delta p_C - \Delta p_E e^{-\frac{W_b}{L_p}}}{e^{\frac{W_b}{L_p}} - e^{-\frac{W_b}{L_p}}} \\ C_2 = \dfrac{\Delta p_E e^{\frac{W_b}{L_p}} - \Delta p_C}{e^{\frac{W_b}{L_p}} - e^{-\frac{W_b}{L_p}}} \end{cases}$$

If we assume that the collector junction is strongly reverse biased and the equilibrium hole concentration p_n is negligible compared with the injected concentration Δp_E, the concentration of excess holes at x_n within the base region becomes:

Eq. (5.9) $$\delta p(x_n) = \Delta p_E \frac{e^{\frac{W_b}{L_p}} e^{-\frac{x_n}{L_p}} - e^{-\frac{W_b}{L_p}} e^{\frac{x_n}{L_p}}}{e^{\frac{W_b}{L_p}} - e^{-\frac{W_b}{L_p}}} \quad \text{(for } \Delta p_C \approx 0\text{)}$$

Having solved for the excess hole distribution in the base region, we can now evaluate the emitter and collector currents from the gradient of the hole concentration at each depletion region edge:

Eq. (5.10) $$I_p(x_n) = -qAD_p \frac{d\delta p(x_n)}{dx_n}$$

This expression evaluated at $x_n = 0$ gives the hole component of the emitter current (i.e. I_{Ep}), and evaluated at $x_n = W_b$ gives the collector current (i.e. I_C):

Eq. (5.11) $$\begin{cases} I_{Ep} = I_p(x_n = 0) = qA\dfrac{D_p}{L_p}(C_2 - C_1) \\ \qquad = qA\dfrac{D_p}{L_p}\left(\Delta p_E \operatorname{ctnh}\dfrac{W_b}{L_p} - \Delta p_C \operatorname{csch}\dfrac{W_b}{L_p} \right) \\ I_C = I_p(x_n = W_b) = qA\dfrac{D_p}{L_p}\left(C_2 e^{-\frac{W_b}{L_p}} - C_1 e^{\frac{W_b}{L_p}} \right) \\ \qquad = qA\dfrac{D_p}{L_p}\left(\Delta p_E \operatorname{csch}\dfrac{W_b}{L_p} - \Delta p_C \operatorname{ctnh}\dfrac{W_b}{L_p} \right) \end{cases}$$

Then I_B is obtained by current summation:

Eq. (5.12) $$I_B = I_E - I_C = qA\frac{D_p}{L_p}\left[(\Delta p_E + \Delta p_C) \tanh\frac{W_b}{2L_p} \right]$$

5.3.4. Current gain

We can now define a few parameters which characterize the amplification mechanism. For simplicity, we will neglect the saturation current at the collector (3 in Fig. 5.5) and recombination in the depletion regions.

We first start by expressing the collector current I_C as a function of the emitter current I_E. The total emitter current I_E has two separate components: a net hole and a net electron diffusion currents (5 in Fig. 5.5), or i_{Ep} and i_{En} respectively:

Eq. (5.13) $$I_E = I_{Ep} + I_{En}$$

An emitter injection efficiency γ can thus be defined as:

Eq. (5.14) $$\gamma = \frac{I_{Ep}}{I_{Ep} + I_{En}}$$

The emitter injection efficiency can be considered as the portion of total emitter current that is due solely to minority carriers being injected into the base (see Fig. 5.7). γ can be closer to unity when the p-type emitter region is highly doped (p^+). It can be shown that the emitter injection efficiency of a p-n-p transistor can also be written in terms of the emitter (L_p^n, p_p) and base material properties (L_n^p, n_n):

Eq. (5.15) $$\gamma = \left[1 + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n} \tanh \frac{W_b}{L_p^n}\right]^{-1} \approx \left[1 + \frac{W_b n_n \mu_n^p}{L_n^p p_p \mu_p^n}\right]^{-1}$$

In this equation we use superscripts to indicate which side of the emitter-base junction is referred to.

The ratio of the collector hole current I_{Cp} to the hole component of the emitter current I_{Ep}, called the base transport factor, is denoted α_T and is given by:

Eq. (5.16) $$\alpha_T = \frac{I_{Cp}}{I_{Ep}} = \frac{\operatorname{csch} \frac{W_b}{L_p}}{\operatorname{ctnh} \frac{W_b}{L_p}} = \operatorname{sech} \frac{W_b}{L_p}$$

This factor reflects the amount of recombination occurring in the base region. Finally, the ratio of the collector current to the total emitter current is called the current transfer ratio and is denoted α_0:

Eq. (5.17) $$\alpha_0 = \alpha_T \gamma = \left[\cosh \frac{W_b}{L_p^n} + \frac{L_p^n n_n \mu_n^p}{L_n^p p_p \mu_p^n} \sinh \frac{W_b}{L_p^n} \right]^{-1}$$

or expressed otherwise as:

Eq. (5.18) $$\frac{I_C}{I_E} = \frac{\alpha_T I_{Ep}}{I_{En} + I_{Ep}} = \alpha_T \gamma \equiv \alpha_0$$

An efficient transistor is one such that $\alpha_T \approx 1$ and $\gamma \approx 1$, and therefore the current transfer ratio, α_0 is close to unity too.

Now, let us consider the relationship between the collector current I_C and the base current I_B. By taking into account all the currents going into and out of the base region, we can express the base current as:

Eq. (5.19) $$I_B = I_{En} + I_{Ep} - I_C = I_{En} + (1 - \alpha_T) I_{Ep} \text{ also change}$$

Using the above relation and the fact that $I_C = \alpha_T I_{Ep}$, we successively get:

Eq. (5.20) $$\begin{aligned} \frac{I_C}{I_B} &= \frac{\alpha_T I_{Ep}}{I_{En} + (1 - \alpha_T) I_{Ep}} \\ &= \frac{\alpha_T [I_{Ep} / (I_{En} + I_{Ep})]}{1 - \alpha_T [I_{Ep} / (I_{En} + I_{Ep})]} \\ &= \frac{\alpha_T \gamma}{1 - \alpha_T \gamma} = \frac{\alpha_0}{1 - \alpha_0} \\ &= \beta \end{aligned}$$

where β is called the base-to-collector current amplification factor. This factor is also commonly seen as h_{FE} in datasheets and the literature. For an efficient transistor, as α_0 is close to unity, the factor β can be large. This means that the collector current is large compared to the base current.

We mentioned earlier that the amplification can be large if the base region is engineered correctly. This can be illustrated by expressing the amplification factor β in terms of two characteristic times: τ_p and τ_t, which are the average hole lifetime in the base and the average transit time of holes from emitter to collector, respectively. To do so, we will also assume a unity emitter injection efficiency ($\gamma = 1$) and a negligible collector saturation current.

Under these conditions, the average hole recombination lifetime τ_p is also the average time that an electron injected from the base contact spends within the base region. Furthermore, the average time that a hole stays within the base region, W_b, is the transit time τ_t given by:

Eq. (5.21) $$\tau_t = \frac{W_b^{\,2}}{2D_h}$$

This transit time can be made much shorter in comparison to the recombination lifetime τ_p by reducing the dimension of the base region, the origin of the Early effect, discussed later. This means in particular that an injected electron can “outlive” an injected hole in the base region. Thus, in order to ensure the overall charge neutrality of the base region, more holes need to be injected from the emitter into the base. In other words, for each injected electron, there will be τ_p/τ_t holes which can traverse the base region before recombination occurs. This in particular means that:

Eq. (5.22) $$\frac{\tau_p}{\tau_t} \approx \frac{I_{Ep}}{I_B}$$

We can therefore qualitatively understand how the amplification process takes place. Since $I_{Ep} - I_B = I_C$ when $\gamma = 1$, we get:

Eq. (5.23) $$I_B\left(\frac{\tau_p}{\tau_t} - 1\right) = I_C$$

Using Eq. (5.20), we get:

Eq. (5.24) $$\frac{I_C}{I_B} = \frac{\tau_p}{\tau_t} - 1 = \beta$$

Since usually $\frac{\tau_p}{\tau_t}$ is generally large, we get:

Eq. (5.25) $\beta = \frac{I_C}{I_B} \approx \frac{\tau_P}{\tau_t}$

Therefore, as the base current I_B can be controlled independently as shown in Fig. 5.2 and is mainly determined by the external circuit parameters, the collector current I_C will be the base current I_B multiplied by the current amplification factor β, which represents current gain.

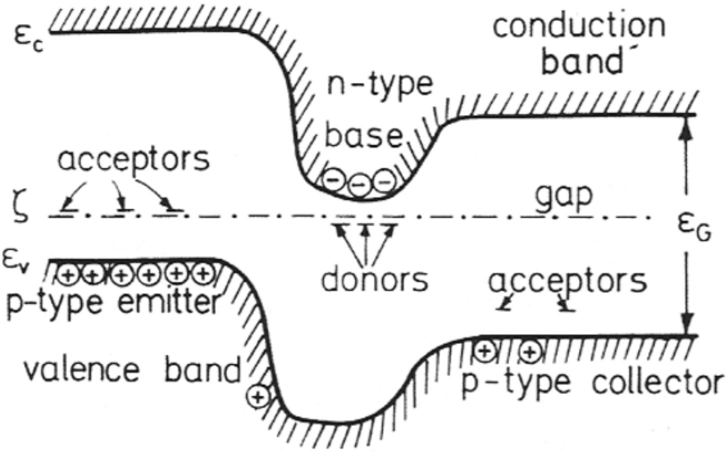


Fig. 5.7. Energy band edges in a p-n-p type transistor at thermal equilibrium. [Semiconductor Physics: An Introduction, 1997, p. 144, Seeger, K., Fig. 5.12. © Springer-Verlag Berlin Heidelberg 1973, 1982, 1985, 1989, 1991 and 1997. With kind permission of Springer Science and Business Media.]

5.3.5. Typical BJT configurations

Four possible modes of operation exist for BJT biasing. The forward active mode is the most commonly used operational mode when using a BJT for amplification purposes (see Fig. 5.7 for the equilibrium state). The three remaining modes are saturation, cutoff, and reverse active modes. The junction biasing is configured as shown in Table 5.1.

Mode of operation	Base-emitter bias	Base-collector bias
Active	Forward	Reverse
Cutoff	Reverse	Reverse
Saturation	Forward	Forward
Reverse Active	Reverse	Forward

Table 5.1. The four modes of operation for a BJT and the corresponding junction biasing for each mode.

The active, cutoff, and saturation modes will be explained in the following sections but for now we will shortly discuss the reverse active mode. The reverse active mode is analogous to forward active in terms the equations applicable to it, except the emitter is replaced with the collector and vice-versa. This mode is not often used due to its poor efficiency arising from the doping configuration and the corresponding depletion widths dimensions.

Fig. 5.8 shows the three common BJT amplifier configurations known as common base, common emitter, and common collector. These configurations can be easily identified by determining which terminal is connected to ground or circuit "common."

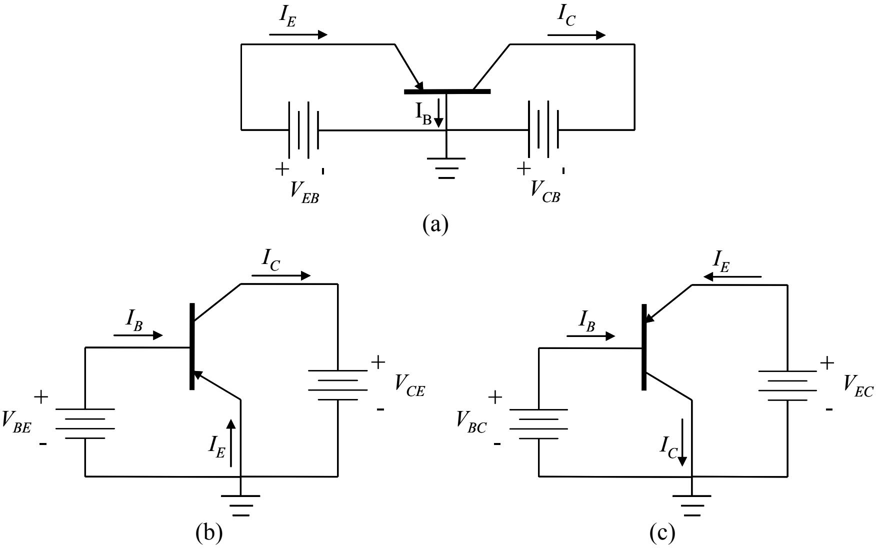


Fig. 5.8. Typical BJT amplifier configurations: (a) common base, (b) common emitter, and (c) common collector.

Consider the case of a common base (CB) configuration. In forward active mode the emitter-base junction is forward biased and the collector-base junction is reverse bias. A typical family of curves for such a configuration is shown in Fig. 5.9.

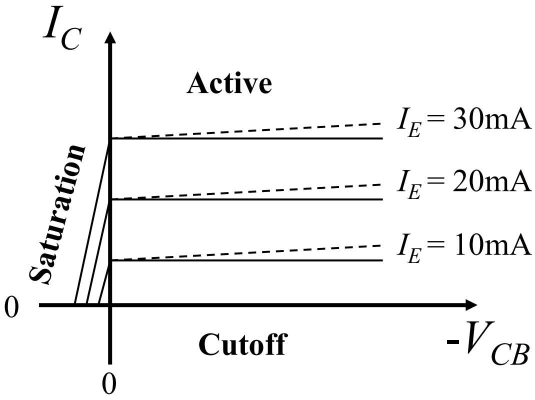


Fig. 5.9. A family of curves demonstrating the dependence of collector current, I_C, on the collector-base voltage, V_{CB}. Curves are shown for a range of emitter currents, I_E. The modes of operation are labeled and the dashed lines indicate the consequence of the Early effect.

In Fig. 5.9 the Early effect is shown with dashed lines. This effect originates from the increasing base-collector depletion width with increasing reverse bias of V_{CB}. This reduces the width of the base region resulting in an increased charge gradient across the base as well as decrease in the recombination probability in the base. The result of the former is increased injection of minority carriers from the emitter and the latter enhances the base transport factor, α_T. If one extrapolates in the positive V_{CB} direction the dashed output curves of Fig. 5.9 they will converge at the Early voltage, V_A. If the base width is much larger than the depletion region extending into the base the Early voltage can be expressed as:

Eq. (5.26) $$V_A \cong \frac{qN_bW_b^2}{\varepsilon_s}$$

where N_b is the base doping.

In the active region of operation the collector-base junction is reverse biased and the emitter-base junction is forward biased. If an emitter current, I_E is allowed to flow then $\sim\alpha I_E$ flows from the collector. In this region the Early effect should be considered but its overall effect is not very significant. If I_E goes to 0 then the collector current is equal to the reverse bias saturation current, I_{CO}, a few nano- to micro-amperes depending upon the material comprising the BJT.

The saturation region refers to the case where both junctions are forward biased. The forward bias behavior of the collector junction dictates the strong dependence of I_C on small changes in V_{CB}.

In the cutoff region of operation both junctions are reverse biased and the collector current is negligible.

The common emitter (CE) configuration is encountered much more often in electronic circuits than the common base configuration and will be considered next. One advantage of the CE configuration, compared to the CB, is that the input current I_B can be much smaller than the output current I_C by nearly β. In the active region it can be shown that:

Eq. (5.27) $$I_C = \beta I_B$$

when the collector leakage current is taken to be negligible and the Early effect is not considered. The active region is the normal region of operation for a CE amplifier. The Early effect has much more dramatic consequences in the case of the CE configuration. A very small change of a fraction of one percent in α due to a decrease in W_b can increase β by tens of percents or more. Proof of this concept is left for the exercises.

The cutoff region is defined as the case when the collector current is equal to the saturation current and the reverse-biased emitter junction's current is zero. If V_{CE} drops below V_{BE} then the collector junction is forward biased and the device is considered to be saturated.

5.3.6. Deviations from the ideal BJT case

As with just about any electrical device, the practical behavior of BJTs deviates, to some respect, from the ideal models presented in the previous sub-section. In this sub-section we will discuss some divergences of BJT behavior from the ideal case. We have already discussed the Early effect so that will not be addressed here but some explanation of base spreading, current crowding, depletion region recombination, and breakdown mechanisms will be offered.

In a *p-n* diode, the occurrence of high injection conditions increases the carrier density of injected carriers to levels similar to that of the collector doping concentration. When a BJT is placed into high injection conditions, by forward biasing the emitter-base junction a reduction in the current gain is realized according to the following relationship with V_{BE}:

Eq. (5.28) $$\beta = \beta_0 \exp\left(-\frac{V_{BE}}{2V_t}\right)$$

Punch through is the case when the low-doped base width is reduced to zero and a short circuit between the *p*-type (or *n*-type) collector and emitter is created upon sufficient V_{CB}. This condition can also arise from very narrow base widths. The result of punch through is a large current through the current and emitter.

Breakdown in BJT devices usually originates from avalanche multiplication. Collector doping levels are typically not high enough to cause direct tunneling, or Zener breakdown. The avalanche process in BJTs is nearly identical to that in a *p-n* junction. In the CE configuration avalanche breakdown is caused by impact ionization. The ionized carriers appear as an increase in the base current which causes even more collector current to flow in a positive feedback fashion. One can show that the breakdown voltage in the CE configuration is:

Eq. (5.29) $$V_{B,CE} = V_{B,CB}(1-\alpha_0)^{1/n}$$

where n is a constant and $V_{B,CB}$ is the common-base breakdown voltage. The common-base breakdown voltage is commonly determined by the open-emitter reverse breakdown voltage of the base-collector *p-n* junction. This breakdown voltage is similar to that for a typical *p-n* junction and, again, is generally due to avalanche breakdown.

5.4. Heterojunction bipolar transistors

In a homojunction BJT, the emitter injection efficiency is limited by the fact that carriers can flow from the base into the emitter region, over the emitter junction barrier, which is reduced by the forward bias. It is necessary to use lightly doped base and heavily doped p^+ emitter for the optimum injection of holes. But this will result in higher base resistance. Degenerate doping can lead to a slight decrease of E_g in the emitter, which will decrease the emitter injection efficiency. For high frequency applications, a heavily doped base and a lightly doped emitter are desirable. There are better ways to accomplish the design instead of doping only, i.e. to use heterojunctions instead of homojunctions. We then talk about a heterojunction bipolar transistor or HBT.

For example, if we use a wider bandgap material for the emitter than the base, then it is possible that for an *n-p-n* transistor, the barrier for electron injection is smaller than the hole barrier. Since carrier injection rate varies exponentially with the barrier height, even a small difference in these two barriers can make a very large difference in the transport of electrons and holes across the emitter junction. Neglecting differences in carrier mobilities and other effects, we can approximate the dependence of carrier injection across the emitter as:

Eq. (5.30) $$\frac{I_n}{I_p} \propto \frac{N_D^E}{N_A^B} e^{\frac{\Delta E_g}{k_b T}}$$

A relatively small value of ΔE_g will have significant effects to the current ratio. This allows us to choose the doping terms for lower base resistance and emitter junction capacitance. In particular, we can choose a heavily doped base to reduce the base resistance and a lightly doped emitter to reduce junction capacitance. However there will usually be spike and notch at heterojunction interface. This can be eliminated by graded interface.

In the following sub-sections, we will describe two of the most widely used heterojunction bipolar transistors: AlGaAs/GaAs and GaInP/GaAs HBTs.

5.4.1. AlGaAs/GaAs HBT

Thanks to the excellent lattice-match between $Al_xGa_{1-x}As$ and GaAs over the entire compositional range, the AlGaAs/GaAs system has been the most widely used system for heterojunction bipolar transistors. Fig. 5.10 shows the cross-section structure of a typical AlGaAs/GaAs HBT.

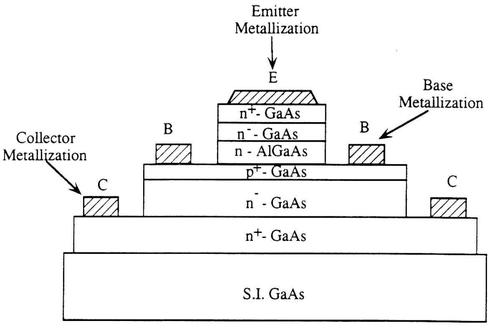


Fig. 5.10. Cross-section structure of an AlGaAs/GaAs heterojunction bipolar transistor. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Fig. 9.22, p. 395. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

In order to avoid DX (unknown defect) center problems, the Al mole fraction x in $Al_xGa_{1-x}As$ is usually kept around 0.25 which results in a conduction-band discontinuity of 0.2 eV and a valence-band discontinuity of 0.1 eV. Due to the large conduction-band discontinuity, the emitter-base junction is usually computationally graded.

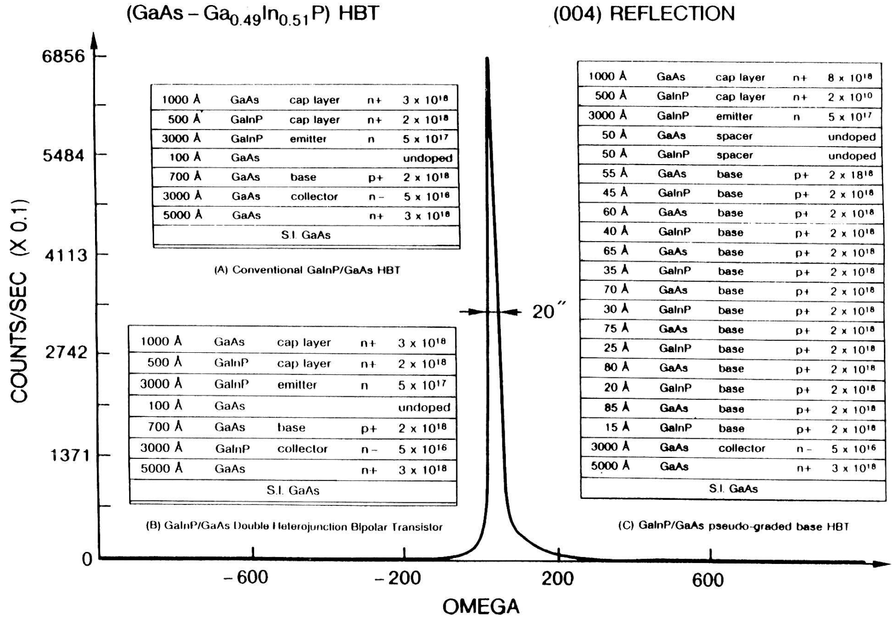
Device isolation is performed through deep ion implantation to make the layers outside the device semi-insulating or by using a mesa structure. By means of composition-selective etches, vias are etched to the base and collector layers to make the corresponding contacts.

In order to obtain good device performance, the contact resistance of the ohmic contacts should be minimized. One of the ways that has been used to reduce the emitter contact resistance is to use lattice-mismatched InGaAs cap layers grown on the emitter. Due to the small metal-semiconductor barrier height good ohmic contacts can be achieved. Commonly used contacts are AuGe/Ni and Ge/Au/Cr. Minimization of the parasitic resistance is also very important in obtaining a good device performance. The base-emitter separation should be a few tenths of a micron. This can be accomplished by using self-aligned techniques ([Nagata et al. 1987] [Hayama et al. 1987] [Chang et al. 1987]). By using a shallow proton implant into the collector region under the base contacts, extrinsic collector doping and therefore the base-collector capacitance can be reduced [Ginoudi et al. 1992].

It is important that the base-emitter *p-n* junction coincides with the heterojunction between AlGaAs and GaAs. Therefore good doping profile and material composition control is required in the growth of the epilayers. Most of the AlGaAs/GaAs HBT research has been done on MBE-grown devices. Since the early 1980s the performance of MOCVD-grown AlGaAs/GaAs HBTs has increased significantly. A f_{max} of 94 GHz and an f_t of 45 GHz have been obtained by Enquist and Hutchby [1989] using a self-aligned structure. One of the difficulties in HBT fabrication is the diffusion of impurities from the heavily doped GaAs base into the AlGaAs emitter at high temperatures during or subsequent to growth. This causes the *p-n* junction to move into the AlGaAs layer and the current gain of the device is reduced due to the reduction in the barrier to hole injection. This problem can be avoided by introducing a thin undated GaAs spacer layer between the base and emitter or by reduction of the growth temperature before the AlGaAs layer is grown. Common *p*-type dopants in MOCVD are magnesium, zinc and carbon. Mg doping shows abnormal memory effects, which requires growth interruptions in order to obtain an abrupt doping profile ([Kuech et al. 1988] [Landgren et al. 1988]). Zn has a large diffusion coefficient and carbon doping needs a low growth temperature both of which are incompatible with the growth of high-quality AlGaAs which requires high temperatures. However, very high base doping levels are possible with carbon doping due to the very low diffusion coefficient of carbon [Ashizawa et al. 1991]. Using carbon doping in the base ($p = 4 \times 10^{19}$ cm^{-3}), Twynam et al. [1991] have reported MOCVD-grown AlGaAs/GaAs microwave HBTs with an f_t of 42 GHz, f_{max} of 117 GHz and a current gain of 50.

5.4.2. GaInP/GaAs HBT

The $Ga_{0.51}In_{0.49}P$/GaAs system has some major advantages over AlGaAs/GaAs. For *n-p-n* heterojunction bipolar transistors, the GaInP/GaAs system has an additional advantage when compared with the widely used AlGaAs/GaAs structure. The valence-band discontinuity in the $Ga_{0.51}In_{0.49}P$/GaAs system is about 0.28 eV and conduction-band discontinuity is 0.2 eV [Biswas et al. 1990]. A large valence-band discontinuity is an exciting property for *n-p-n* HBTs. In the AlGaAs/GaAs system, the same amount of valence-band discontinuity requires that the Al mole fraction be about 0.6, in which case there would be a very large conduction-band spike at the emitter-base junction together with an indirect-gap emitter, neither being acceptable. In the AlGaAs/GaAs system, about 60 per cent of the energy gap difference occurs in the conduction band and the emitter-base junction of the device is usually graded to eliminate the conduction-band spike which decreases the emitter injection efficiency and increases the emitter switch-on voltage. However, theoretical investigations [Das and Lundstrom 1988] have shown that grading of the emitter-base junction increases the recombination in the emitter-base junction and therefore the current gain may not be increased considerably by junction grading. Because of the relatively small conduction-band discontinuity and large valence-band discontinuity of $Ga_{0.51}In_{0.49}P$/GaAs, it can be estimated that the current gain of *n-p-n* HBTs based on this material system will be significantly higher than that of AlGaAs/GaAs HBTs. Modry and Kroemer [1985] have reported a GaInP/GaAs HBT grown by MBE. The current gain was low at small current densities suggesting a high recombination rate at the emitter-base junction due to a large number of defects at the heterojunction interface. A maximum current gain of 30 was obtained at 3000 A·cm^{-2}. Later, MOCVD and chemical beam epitaxy grown GaInP/GaAs HBTs with better performances were reported ([Kobayashi et al. 1989] [Razeghi et al. 1990] [Alexandre et al. 1990] [Bachem et al. 1992]). In addition, Razeghi et al. [1990] reported a current gain of 400 for a low-pressure MOCVD-grown GaInP/GaAs HBT. Three different HBT structures were grown: (i) conventional, (ii) double heterojunction, (iii) pseudo-graded base. The details of collector, base and emitter thicknesses, carrier concentrations and a typical x-ray diffraction pattern for the structures around the (400) reflection peak are shown in Fig. 5.11.



1000 Å	GaAs	cap layer	n+	3 x 10^{18}
500 Å	GaInP	cap layer	n+	2 x 10^{18}
3000 Å	GaInP	emitter	n	5 x 10^{17}
100 Å	GaAs			undoped
700 Å	GaAs	base	p+	2 x 10^{18}
3000 Å	GaAs	collector	n –	5 x 10^{16}
5000 Å	GaAs		n+	3 x 10^{18}
S.I. GaAs				

(A) Conventional GaInP/GaAs HBT

1000 Å	GaAs	cap layer	n+	3 x 10^{18}
500 Å	GaInP	cap layer	n+	2 x 10^{18}
3000 Å	GaInP	emitter	n	5 x 10^{17}
100 Å	GaAs			undoped
700 Å	GaAs	base	p+	2 x 10^{18}
3000 Å	GaInP	collector	n –	5 x 10^{16}
5000 Å	GaAs		n+	3 x 10^{18}
S.I. GaAs				

(B) GaInP/GaAs Double Heterojunction Bipolar Transistor

1000 Å	GaAs	cap layer	n+	8 x 10^{18}
500 Å	GaInP	cap layer	n+	2 x 10^{10}
3000 Å	GaInP	emitter	n	5 x 10^{17}
50 Å	GaAs	spacer		undoped
50 Å	GaInP	spacer		undoped
55 Å	GaAs	base	p+	2 x 18^{18}
45 Å	GaInP	base	p+	2 x 10^{18}
60 Å	GaAs	base	p+	2 x 10^{18}
40 Å	GaInP	base	p+	2 x 10^{18}
65 Å	GaAs	base	p+	2 x 10^{18}
35 Å	GaInP	base	p+	2 x 10^{18}
70 Å	GaAs	base	p+	2 x 10^{18}
30 Å	GaInP	base	p+	2 x 10^{18}
75 Å	GaAs	base	p+	2 x 10^{18}
25 Å	GaInP	base	p+	2 x 10^{18}
80 Å	GaAs	base	p+	2 x 10^{18}
20 Å	GaInP	base	p+	2 x 10^{18}
85 Å	GaAs	base	p+	2 x 10^{18}
15 Å	GaInP	base	p+	2 x 10^{18}
3000 Å	GaAs	collector	n –	5 x 10^{16}
5000 Å	GaAs		n+	3 x 10^{18}
S.I. GaAs				

(C) GaInP/GaAs pseudo-graded base HBT

Fig. 5.11. X-ray diffraction spectrum of a GaInP/GaAs heterojunction bipolar transistor. (A) conventional HBT, (B) double heterojunction HBT, and (C) pseudo-graded base HBT, with the layer thicknesses and carrier concentrations of the collector, base and emitter. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Fig. 9.23, p. 398. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

The diffraction peak is very intense and has a full width at half maximum of 20 seconds, demonstrating that GaInP is perfectly lattice-matched to GaAs and the pseudo-graded base has excellent crystallographic properties, which is necessary to allow optimal transport properties of the injected minority carriers. Fig. 5.12 shows the doping profile demonstrating an abrupt and perfectly controlled transition from emitter to base and from base to the collector.

The device structure was a conventional mesa type. NH_4:H_2O_2:H_2O (10:4:500) and HCI:H_3PO_4 (1:1) were used to etch GaAs and GaInP respectively. The emitter and collector contacts were defined by depositing and annealing Ge/AuNi/Au. The base contact was defined by the deposition and annealing of Zn/Au. Fig. 5.13 shows the emitter-grounded current-voltage characteristic of the device which exhibited a current gain of 400 at 20 mA.

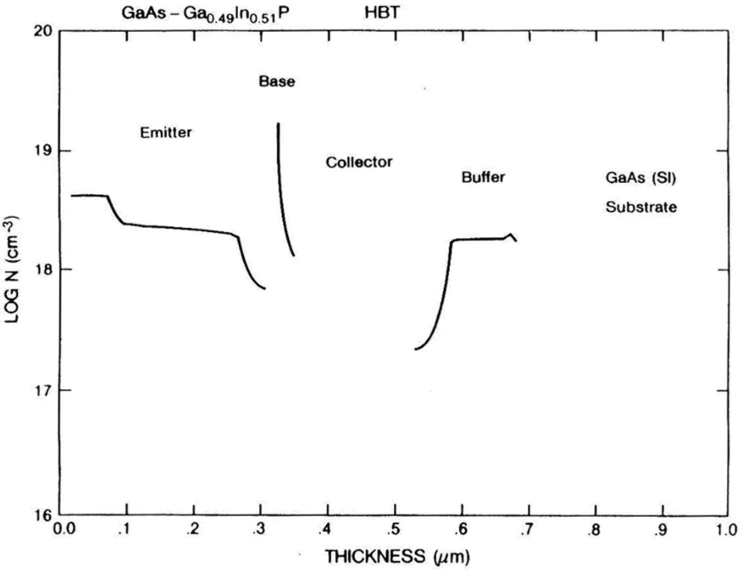


Fig. 5.12. Example of doping profile of a GaInP/GaAs heterojunction bipolar transistor. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Fig. 9.24, p. 399. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

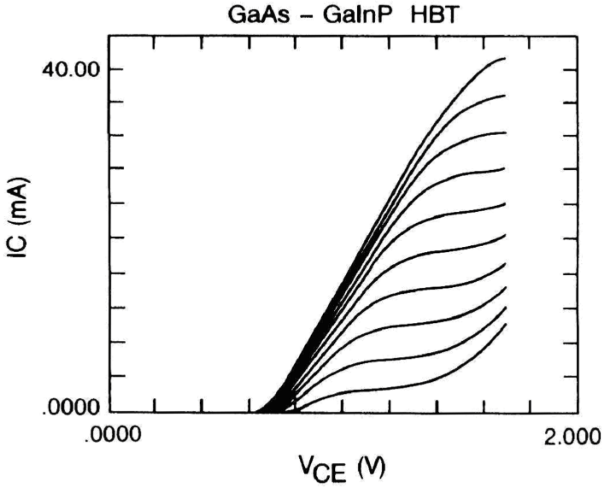


Fig. 5.13. Emitter-grounded current-voltage characteristic of the conventional GaInP/GaAs heterojunction bipolar transistor shown in (A) of Fig. 5.11. [Copyright © 1995 From The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications. Fig. 9.25, p. 399. Reproduced by permission of Routledge/Taylor & Francis Group, LLC.]

5.5. Field effect transistors

A field effect transistor or FET is a three terminal device in which the current flow between two terminals can be controlled by the third terminal. However, unlike a bipolar junction transistor, the control is through the voltage, not current, of the third terminal. There are several types of FETs depending on the junction of the controlling terminal or “gate”. The first type is a junction FET or JFET, where the gate junction is a simple *p*-*n* junction. If this junction is replaced with a metal-semiconductor Schottky contact, the device is called a metal-semiconductor FET or MESFET. Also, if an insulator is placed between the metal and the semiconductor, the device is a metal-insulator-semiconductor FET or MISFET. Oxides are the common insulator, and the devices based on oxide insulators are metal-oxide-semiconductor FET or MOSFET.

5.5.1. JFETs

The operation of FET is based on the change of the thickness of a conducting layer or channel, and hence the current flow through it. Fig. 5.14(a) shows the schematic diagram of a JFET.

The device is made from an *n*-type channel sandwiched between two *p*-type “gate” layers. The two ends of the channel are attached to metal contacts and are named drain and source. There are two depleted layers that are naturally formed between the *n*-type channel and the *p*-type gates. Under a zero bias, the thicknesses of the depleted layers are constant. However, if current *I* passes through the channel, the resistance of the channel results in a voltage gradient across it (Fig. 5.14(b)). This means that the voltage between the gate and the channel is higher at the drain compared to the source, and hence the thickness of the depleted layer is higher at the drain accordingly. The thickness of the depleted layer increases for higher gate biases, and thus higher channel currents, and at some point the depleted layers from both sides of the channel reach together. This situation is called pinch-off (Fig. 5.14(c)) and it prevents a further increase of the channel current even if a higher voltage is applied between the drain and the source.

After the initial pinch-off and upon further gate bias, the pinch-off point near the drain moves towards the source. The *n*-type channel from the source to the pinch-off point dominates the resistance of the electron flow through the *n*-type channel until the electrons are quickly swept across the highly resistive depletion region by the large electric field.

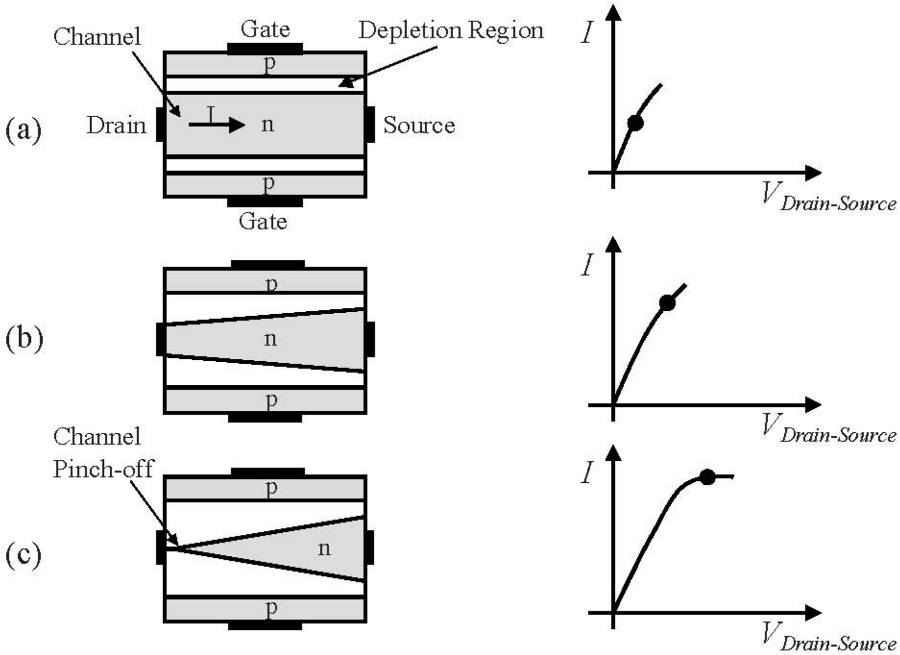

Channel
Gate
Depletion Region
(a)
Drain
I
n
Source
p
p
Gate
(b)
p
n
p
Channel
Pinch-off
(c)
p
n
p
I
$V_{Drain\text{-}Source}$
I
$V_{Drain\text{-}Source}$
I
$V_{Drain\text{-}Source}$

Fig. 5.14. Schematic diagrams of a junction field effect transistor under different operation conditions. (a) With a low bias between the drain and the source, the thicknesses of the depleted layers are nearly constant. (b) When a larger bias is applied and a higher current results across the channel, the depleted layers get thicker near the drain. (c) At some point, the depleted layers from both sides of the channel reach, resulting in the pinch-off condition when no further increase in the channel current is possible even if a higher bias is applied.

5.5.2. JFET gate control

A negative bias of the gate can simply increase the thickness of the depleted layer, and change the effective thickness of the channel. This means that the conductance of the channel can be reduced with a negative bias on the gate. More importantly, the pinch-off effect happens at a lower drain-source current. Fig. 5.15 shows the current-voltage relationship of the drain-source with different gate voltages.

Note that if the drain-source voltage is higher than the pinch-off condition, the drain-source current only depends on the gate voltage. Therefore, the device behaves as a current source that is controlled by the gate voltage. Such a characteristic is very useful in the design of AC amplifiers.

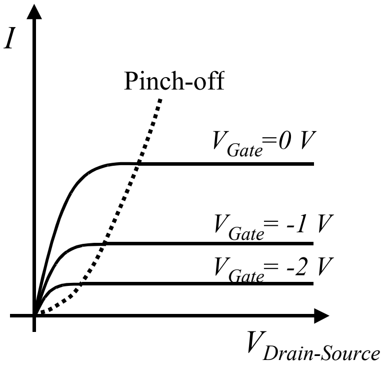


Fig. 5.15. Current-voltage relationship between the drain and the source of a field effect transistor as a function of gate voltage. The dotted curve shows the characteristic points where the pinch-off occurs.

The pinch-off voltage of the device can be calculated using our knowledge about the gate-channel *p-n* junctions. Assuming that the gates are heavily doped (p^+) and the built-in voltage of the junction V_0, is negligible compared to the gate-drain voltage V_{GD}, the depletion layer thickness is:

Eq. (5.31) $$W = \left(\frac{2\varepsilon(-V_{GD})}{qN_D} \right)^{1/2}$$

where ε is the permittivity of the semiconductor and N_D is the donor concentration in the channel. Now at the gate-drain voltage that pinch-off happens or $-V_p$ the depletion thickness is equal to the channel width a, we have:

Eq. (5.32) $$a = \left(\frac{2\varepsilon(V_p)}{qN_D} \right)^{1/2} \rightarrow V_p = \frac{qa^2 N_D}{2\varepsilon}$$

5.5.3. JFET current-voltage characteristics

Now we are in a position to calculate the current-voltage characteristic of a JFET. Fig. 5.16 shows a simplified diagram of the device.

Considering the symmetry of the device, the half of the channel width is called a, and the depletion layer from one side is $W(x)$ where the origin of the coordinate x is placed at the drain. The conducting part of the channel is $a-W(x) = h(x)$. Now if the width of the channel is Z, the total area of the channel at position x is:

Eq. (5.33) $A = 2h(x)Z$

and assuming that the resistivity of the channel is ρ, the resistance of the channel over a differential thickness *dx* is:

Eq. (5.34) $R(x) = \rho\frac{dx}{A} = \rho\frac{dx}{2h(x)Z}$

and the drain-source current *I*, is simply the voltage drop dV_x over the differential distance *dx* divided by the resistance *R(x)*:

Eq. (5.35) $I_{DS} = \frac{-dV_x}{R(x)} = -\frac{2Zh(x)dV_x}{\rho dx}$

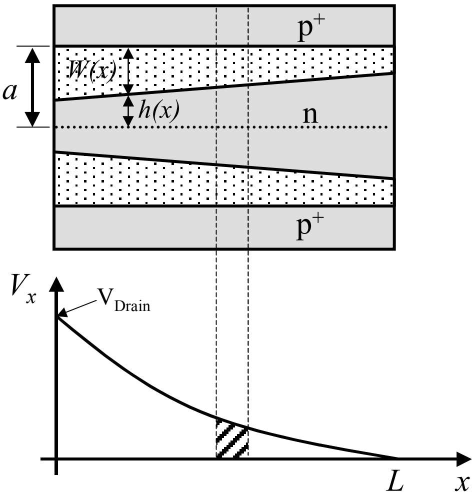


Fig. 5.16. Schematic diagram of a junction field effect transistor, showing the depletion layer width W(x) as a function of the distance from the drain x.

Now *h(x)* can be replaced with:

Eq. (5.36) $h(x) = a - W(x) = a - \left(\frac{2\varepsilon(V_x - V_G)}{qN_D}\right) = a\left[1 - \left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]$

Inserting Eq. (5.36) into Eq. (5.35), we have:

Eq. (5.37) $$I_{DS} = -\frac{2ZdV_x}{\rho dx} a\left[1-\left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]$$

and now we can separate I and dV_x as:

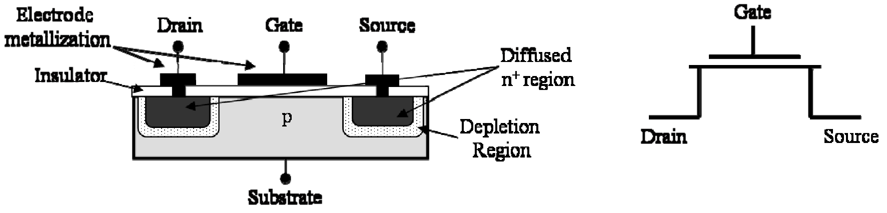
Eq. (5.38) $$I_{DS}dx = -\frac{2Z}{\rho} a\left[1-\left(\frac{V_x - V_G}{V_p}\right)^{1/2}\right]dV_x$$

Integrating from both sides yield:

Eq. (5.39) $$I_{DS} = G_0 V_p\left[\frac{V_D}{V_p}+\frac{2}{3}\left(-\frac{V_G}{V_p}\right)^{3/2}-\frac{2}{3}\left(\frac{V_D - V_G}{V_p}\right)^{3/2}\right]$$

5.5.4. MOSFETs

We will now take a look at the metal oxide semiconductor (MOS)-based FETs. The operation of MOSFETs is based on the effect of an electric field penetrating into the conductive channel between two highly-doped contact regions for the source and drain. The basic structure of an n-type MOSFET is shown in Fig. 5.17 along with a typical circuit schematic representation of a MOSFET.


Electrode metallization
Drain
Gate
Source
Insulator
p
Diffused n⁺ region
Depletion Region
Substrate
Gate
Drain
Source

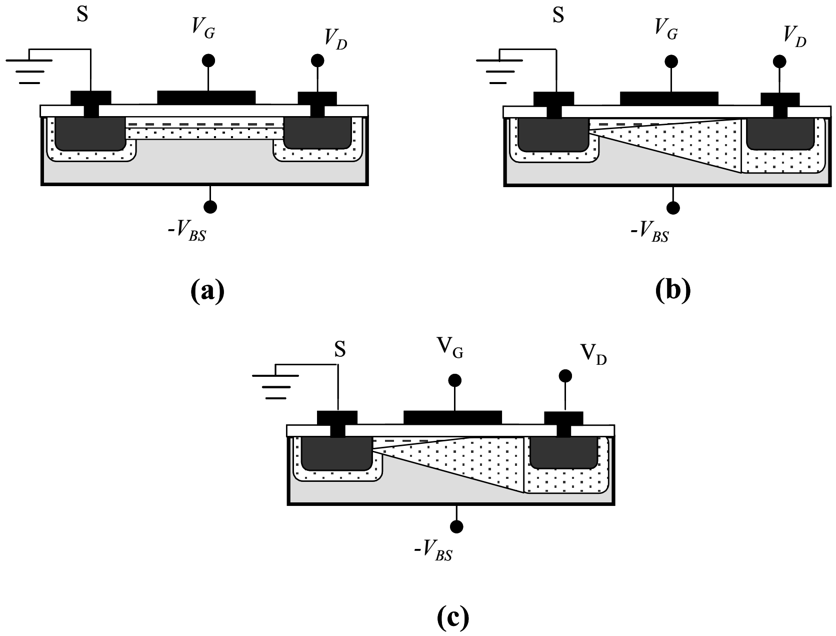
(a) (b)

Fig. 5.17. (a) A schematic depiction of an NMOS FET. (b) The corresponding typical circuit schematic symbol.

The highly-doped n^+ regions are typically diffused into the p-type substrate and act as contact regions for the source and drain metal electrodes. The gate is electrically insulated from the substrate by an insulator, an oxide in the case MOSFETs.

By applying a positive bias to the gate electrode an electric field will extend into the substrate and deplete holes from the region directly under the gate electrode. As the gate bias is increased past some threshold value, V_{th}, an *n*-channel inversion layer forms under the gate and a conductive *n*-type current path between the n^+ source and drain is created.

Based on the biasing configuration there are two basic modes of operation for MOSFET devices; linear and saturation. The situation explained in the previous paragraph describes the linear mode of operation where an increase of the drain-source voltage, V_{DS}, will result in a linear increase in the drain current, I_D, depending upon the resistance of the channel. Similar to the case of the JFET, if V_{DS} is increased further the channel begins to pinch-off and the drain current saturates. This mode of operation is called the saturation region. These two modes of operation are depicted in Fig. 5.18(a) and (b).



(a) **(b)**

(c)

Fig. 5.18. (a) In the linear mode of operation of a MOSFET, where $V_{DS} < V_{GS} - V_{th}$, the drain current increases linearly with the drain voltage, V_D. (b) In the saturation mode of operation, where $V_{DS} > V_{GS} - V_{th}$, pinch-off occurs and the drain current saturates. (c) Upon further increase of V_{DS} the channel length is shortened and drain current increases even above that for the saturation case.

The relationship between drain current and V_{DS} for both modes of operation can be shown to be:

Eq. (5.40) $$I_D = \mu C_{ox}\frac{W}{L}\left[(V_{GS}-V_{th})V_{DS}-\frac{V_{DS}^2}{2}\right]$$

when $V_{DS} < (V_{GS}-V_{th})$ and the MOSFET is operating in the linear region and:

Eq. (5.41) $$I_D = \mu C_{ox}\frac{W}{L}\frac{(V_{GS}-V_{th})^2}{2}$$

when $V_{DS} > (V_{GS}-V_{th})$ and the MOSFET is in its saturation mode of operation, μ is the channel mobility, C_{ox} is the gate oxide capacitance, W is the channel width (into the page), L is the channel length, V_{GS} is the gate-source voltage, and V_{th} is the threshold voltage (~0.5–1 V for silicon-based devices).

5.5.5. Deviations from the ideal MOSFET case

Similar to BJT devices, the simplified MOSFET principles of operation and corresponding relationships do not fully explain the practical behavior of actual MOSFET devices. In this sub-section we will discuss velocity saturation, channel-length modulation, and insulator breakdown.

In an effort to keep pace with Moore's law, MOSFET transistor sizes have continuously followed a trend of decreasing minimum feature size. This scaling reduces parameters such as oxide thickness, gate length, transit time, current, power consumption, voltage, etc. Such scaling has strong benefits in terms of economics and performance but can cause a variety of short channel effects that complicate transistor and highly-integrated circuit design.

One of the most significant effects encountered with decreasing channel lengths is velocity saturation. At low electric fields, a linear relationship between carrier velocity and electric field strength is observed, even for short channels. But at higher field strengths, the carrier velocities begin to saturate and are on the order of the thermal velocity. The velocity of carriers under a high electric field saturates due to increased optical phonon emission. An approximation of this effect is given by the following analytical expression:

Eq. (5.42) $$v_d = \frac{\mu_n E}{1+E/E_c}$$

where v_d is the carrier velocity, μ_n is the low-field mobility, E is the electric field, and E_c is the critical field value.

As discussed in the previous sub-section, there is a drain-source voltage, V_p, above which the conduction channel begins to pinch off. Increasing V_{DS} to a value greater than V_p, moves the pinch-off point towards the source, as shown in Fig. 5.18(c). This effectively reduces the channel length and increases the drain current, I_D. This effect can be mitigated by increasing the substrate carrier concentration. If V_{DS} is increased even further, punch-through can occur and the drain and source are effectively short circuited, similarly to the punch-through case of bipolar transistors.

An additional limitation on applied terminal voltages pertains to the gate voltage. Excessive gate voltages can cause the gate dielectric to catastrophically breakdown. This voltage is dependent upon the dielectric type and thickness, but is typically around 25~50 V.

5.6. Application specific transistors

A brief summary of a few application specific transistor types will now be given. We will discuss single electron transistors, power transistors, and high electron mobility transistors.

Single electron transistors (SET) are metal-insulator-metal (MIM)-based devices that operate on the concept of electron tunneling. By placing two such MIM junctions in series with a gate capacitor connected to a third electrode between the two MIM junctions the SET device structure is realized. By increasing the voltage on the gate capacitor, electrons tunnel more quickly and the current through the device is increased. This makes an SET similar to a MOSFET but on a much smaller scale. If the gate capacitor is made even smaller, fewer electrons are involved in the tunneling current and quantization effects become more prominent.

Power transistors exist for both bipolar and MOSFET transistor types. Bipolar transistors are more traditionally used due to their robust ability to withstand high currents. This is generally due to the larger active area of these devices which allow for low current densities but high, up to 1000 A, currents. Silicon BJTs are most commonly encountered due to their relatively low cost of manufacturing but silicon carbide (SiC) is fundamentally capable of higher breakdown voltages and better thermal conductivity than silicon, but at a higher cost. Other bipolar power transistors include Darlington transistors, thyristors, insulated gate bipolar transistors, and triacs.

High electron mobility transistors (HEMTs) are used in very low noise amplifiers and very high frequency applications such as microwave radio frequency, space communications, radio-based telescopes, and digital broadcasting systems. They are constructed of III-V compound semiconductor materials such as GaAs/AlGaAs, for example. In HEMTs, the heterojunction formed creates a two-dimensional electron gas (2DEG) which confines electrons to a very thin, high-mobility conduction layer resulting in a very low channel resistivity. By applying a gate voltage the conductivity of this channel is changed, effecting the current flow through the device, giving it its transistor-like behavior.

5.7. Summary

In this Chapter, we have described the general principles for electrical amplification and switching. We then modeled the amplification mechanisms, the charge distribution and transport in bipolar junction transistors. The advantages of heterojunction bipolar transistors have been discussed and illustrated with transistors based on AlGaAs/GaAs or GaInP/GaAs. Finally, the principles and electrical properties of field effect transistors were presented.

References

Alexandre, F., Benchimol, J.L., Dangla, J., Dubon-Chevallier, C., and Amarger, V., "Heavily doped based GaInP/GaAs heterojunction bipolar-transistor grown by chemical beam epitaxy," *Electronics Letters.* 26, pp. 1753-1755, 1990.

Ashizawa, Y., Noda, T., Morizuka, K., Asaka, M., and Obara, M., "LPMOCVD growth of C-doped GaAs-layers and AlGaAs/GaAs heterojunction bipolar-transistors," *Journal of Crystal Growth* 107, pp. 903-908, 1991.

Bachem, K.H., Lauterbach, Th., Maier, M., Pletschen, W., and Winkler, K., "MOVPE growth, technology and characterization of $Ga_{0.5}In_{0.5}P$/GaAs heterojunction bipolar-transistors," *Gallium Arsenide and Related Compounds* (*Institute of Physics Conference Series* 120), ed. G.B. Stringfellow, Institute of Physics, Bristol, pp. 293-298, 1992.

Biswas, D., Debhar, N., Bhattacharya, P., Razeghi, M., Defour, M., and Omnes, F., "Conduction-band and valence-band offsets in GaAs/$Ga_{0.51}In_{0.49}P$ single quantum-wells grown by metalorganic chemical vapor-deposition," *Applied Physics Letters* 56, pp. 833-835, 1990.

Chang, M.F., Asbeck, P.M., Wang, K.C., Sullivan, C.J., Sheng, N.H., Higgins J.A., and Miller, D.L., "AlGaAs/GaAs heterojunction bipolar-transistors fabricated using a self-aligned dual lift-off process," *IEEE Electron Device Letters* 8, pp. 303-305, 1987.

Das, A. and Lundstrom, M.S., "Numerical study of emitter-base junction design for AlGaAs GaAs heterojunction bipolar-transistors," *IEEE Transactions on Electron Devices* 35, pp. 863-870, 1988.

Enquist, P.M. and Hutchby, J.A., "High-frequency performance of MOVPE npn AlGaAs/GaAs heterojunction bipolar-transistors," *Electronics Letters* 25, pp. 1124-1125, 1989.

Ginoudi, A., Paloura, E.C., Kostandinidis, G., Kiriakidis, G., Maurel, Ph., Garcia, J.C., and Christou, A., "Low-temperature DC characteristics of S-doped and Si-doped $Ga_{0.51}In_{0.49}P$/GaAs high electron-mobility transistors grown by metalorganic molecular-beam epitaxy," *Applied Physics Letters* 60, pp. 3162-3164, 1992.

Hayama, N., Okamoto, A., Madihian, M., and Honjo, K., "Submicrometer fully self-aligned AlGaAs/GaAs heterojunction bipolar-transistor," *IEEE Electron Devices Letters* 8, pp. 246-248, 1987.

Kobayashi, T., Taira, K., Nakamura, F., and Kawai, H., "Band lineup for a GaInP/GaAs heterojunction measured by high-gain npn heterojunction bipolar-transistor grown by metalorganic chemical vapor-deposition," *Journal of Applied Physics* 65, pp. 4898-4902, 1989.

Kuech, T.F., Wang, P.J., Tischler, M.A., Potenski, R., Scilla, G.J., and Cardone, F., "The control and modeling of doping profiles and transients in MOVPE growth," *Journal of Crystal Growth* 93, pp. 624-630, 1988.

Landgren, G., Rask, M., Anderson, S.G., and Lundberg, A., "Abrupt Mg doping profiles in GaAs grown by metalorganic vapor-phase epitaxy," *Journal of Crystal Growth* 93, pp. 646-649, 1988.

Mondry, M.J. and Kroemer, H., "Heterojunction bipolar-transistor using a (Ga,In)P emitter on a GaAs base, grown by molecular-beam epitaxy," *IEEE Electron Device Letters* 6, pp. 175-177, 1985.

Nagata, K., Nakajima, O., Nittono, T., Ito, H., and Ishibashi, T., "Self-aligned AlGaAs/GaAs HBT with InGaAs emitter cap layer," *Electronics Letters* 23, pp. 64-65, 1987.

Razeghi, M., Omnes, F., Defour, M., Maurel, Ph., Hu, J., Wolk, E., and Pavlidis, D., "High-performance GaAs GaInP heterostructure bipolar-transistors grown by low-pressure metalorganic-chemical vapor-deposition," *Semiconductor Science and Technology* 5, pp. 278-280, 1990.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for photonic and electronic device applications*, Institute of Physics, Bristol, UK, 1995.

Seeger, K., *Semiconductor Physics: An Introduction*, Springer-Verlag, New York, 1997.

Twynam, J.K., Sato, H., and Kinosada, T., "High-performance carbon-doped base GaAs AlGaAs heterojunction bipolar-transistor grown by MOCVD," *Electronics Letters* 27, pp. 141-142, 1991.

Problems

1. Explain why BJTs are considered minority carrier devices and FETs are majority carrier devices?

2. Why is an *n-p-n* BJT used for high speed applications rather than a *p-n-p* BJT?

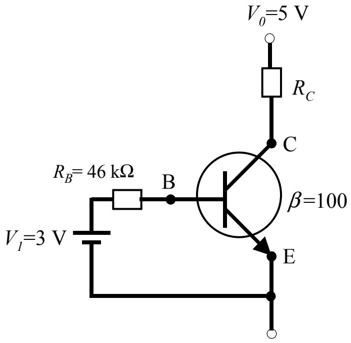
3. What is the typical difference in doping between the emitter and collector in a BJT and why?

4. What is the origin of the Early voltage in a BJT?

5. Thoroughly explain why a BJT performs so poorly in reverse active mode.

6. Summarize the four different modes of operation for a BJT.

7. Consider the electrical circuit shown below, which is designed to deliver a constant current through the collector. Estimate the range of values for the resistance R_C in order to keep this constant current source working properly. Assume that the collector to emitter voltage should be greater than 2 V and that V_{BE} = 0.7 V.



8. Taking into account carrier recombination in the depletion region the current transfer ratio can be expanded to:

$$\alpha_0 = \alpha_T \gamma \delta$$

where δ is the depletion region recombination factor.

(a) Calculate the collector current, base current, DC current gain, and the current transfer ration of a BJT with an I_E of 2.5 mA and the following performance parameters:

α_T= 0.998 (base transport factor)
γ= 0.999 (emitter efficiency)
δ= 0.997 (depletion recombination factor)

(b) Due to the Early effect the value of α_0 in part a is reduced by 0.004. By what percent does this affect the DC current gain of the BJT?

9. Consider a symmetrical p^+-n-p^+ Si bipolar junction transistor with the following properties: $A = 10^{-4}$ cm^2, $W_b = 0.8$ µm, $N_A = 3\times10^{17}$ cm^{-3}, and $N_D = 3\times10^{15}$ cm^{-3}.
   The characteristics of the emitter material are:
   $\tau_n = 0.15$ µs
   $\mu_p = 300$ cm^2/V·s
   $\mu_n = 800$ cm^2/V·s.
   The characteristics of the base material are:
   $\tau_p = 8$ µs
   $\mu_n = 1500$ cm^2/V·s
   $\mu_p = 500$ cm^2/V·s
   Calculate the saturation current on the collector side.

10. Calculate the pinch-off voltage for a silicon nMOSFET with a channel half width of 1.5 µm, and a donor concentration of 2.0×10^{15} cm^{-3}, $\varepsilon_s = 12$.

11. Calculate I_D for an enhancement-mode nMOSFET with a length of $L = 1.3$ µm, width of $W = 15$ µm, an oxide thickness of $t_{\mathrm{ox}} = 25$ nm, and a threshold voltage $V_{\mathrm{T}} = 0.75$ V. The drain-source voltage $\mathrm{V_{DS}} = 5$ V and the gate voltage $\mathrm{V_{GS}} = 3.5$ V. Assume zero substrate bias and a mobility of 350 cm^2/V·s.

6. Semiconductor Lasers

6.1. **Introduction**
6.2. **Types of lasers**
6.3. **General laser theory**
  6.3.1. Stimulated emission
  6.3.2. Resonant cavity
  6.3.3. Waveguides
  6.3.4. Laser propagation and beam divergence
  6.3.5. Waveguide design considerations
6.4. **Ruby laser**
6.5. **Semiconductor lasers**
  6.5.1. Population inversion
  6.5.2. Threshold condition and output power
  6.5.3. Linewidth of semiconductor laser diodes
  6.5.4. Homojunction lasers
  6.5.5. Heterojunction lasers
  6.5.6 Device fabrication
  6.5.7. Separate confinement and quantum well lasers
  6.5.8. Laser packaging
  6.5.9. Distributed feedback lasers
  6.5.10. Material choices for common interband lasers
  6.5.11. Interband lasers
  6.5.12. Quantum cascade lasers
  6.5.13. Type II lasers
  6.5.14. Vertical cavity surface emitting lasers
  6.5.15. Low-dimensional lasers
  6.5.16. Raman lasers
6.6. **Summary**

6.1. Introduction

The word “laser” is an acronym for “light amplification by stimulated emission of radiation”. The principles of lasers were understood at the end

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_6,
© Springer Science+Business Media, LLC 2010

of 1950’s [Schawlow et al. 1958]. The first working laser was built by Maiman in 1960, and used a ruby crystal optically pumped by a flash lamp. The Nobel Prize for fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the laser principle was awarded in 1964 to N.G. Basov, A.M. Prokhorov, and C.H. Townes.

The use of carrier injection across a *p-n* junction for stimulated emission from semiconductors was suggested as early as 1961. The stimulation emission itself was observed in a GaAs *p-n* junction one year later [Holonyak et al. 1962], and then the first semiconductor lasers were fabricated. The light emitted from a laser can be a continuous beam of low or medium power, or it can consist of short bursts of intense light delivering millions of watts.

This Chapter will first review the fundamental mechanisms of a laser. It will then describe the first laser, which was realized using a ruby crystal and subsequently focus on more sophisticated semiconductor lasers.

6.2. Types of lasers

Over the past forty years, scientists have investigated and developed many types of lasers. These lasers fall into several broad categories, as categorized in Fig. 6.1.

Solid Laser
Ruby
Nd:YAG
Ti:Sapphire

Gas Laser
HeNe
Argon-Ion
CO_2

Liquid (Dye) Laser
Polyphenyl
Rhodamine
Oxazine

Semiconductor Laser
III-V
II-VI
IV-VI

Fig. 6.1. Different types of lasers with several examples for each type.

Solid lasers are typically crystals that are doped with specific impurities, which introduce energy levels in the band structure of the crystal. These energy levels determine the energy (or wavelength) of the light emitted by the laser. Solid lasers need to be optically pumped in order to emit light, i.e. the energy needed to make the laser emit light is provided by illuminating the crystal with intense light. This is done typically with an incoherent white flash lamp, though many commercial lasers are now incorporating semiconductor lasers tuned to the optimum absorption frequency for higher efficiency. The power conversion efficiency of a laser is the ratio of the output power it emits to the total power used. Typical power conversion efficiency for a solid laser ranges from 0.1~5%.

Gas lasers are very similar to solid lasers. However, instead of impurities, the energy of the light emitted depends on the gas mixture used. The gas mixture is excited by an electrical discharge. Due to a low absorption efficiency, high voltage discharge (typically 2~4 kV) is used to transfer energy to the gas mixture. The mixture normally consists of an inert gas that absorbs the discharge energy and transfers it to an active gas atom, whose allowed energy levels determine the emission energy. Due to a large variety of gas mixtures, the power conversion efficiency of these lasers ranges from 0.01~15%.

Liquid lasers are typically based on organic dyes dissolved in a solvent. These liquids exhibit groups of many closely spaced energy levels, which can provide a significant amount of emission wavelength tuning (around 90 nm in the visible range). Dye lasers are optically pumped, either with a flash lamp or another laser. The power conversion efficiency of this type of laser can be as high as 20%.

The last type of laser relevant to this discussion is the semiconductor lasers. When electrons in a direct bandgap material relax from the conduction band to the valence band, a photon i.e. light can be emitted. The wavelength of the emitted light can be changed widely through the use of various semiconductor materials with different bandgaps. Semiconductor lasers can be based on III-V, II-VI, and IV-VI compound semiconductors. This Chapter will deal primarily with III-V semiconductor lasers.

Although semiconductor lasers can emit light when optically pumped, they provide a significant advantage of the other types of lasers in that they can also be pumped electrically. Further, unlike gas lasers, the voltage requirements are minimal (1~2 V), which enables semiconductor lasers to be operated with the aid of only a simple battery. Semiconductor lasers are very efficient (up to 80% power efficiency) and come in very small packages, similar to electrical transistors.

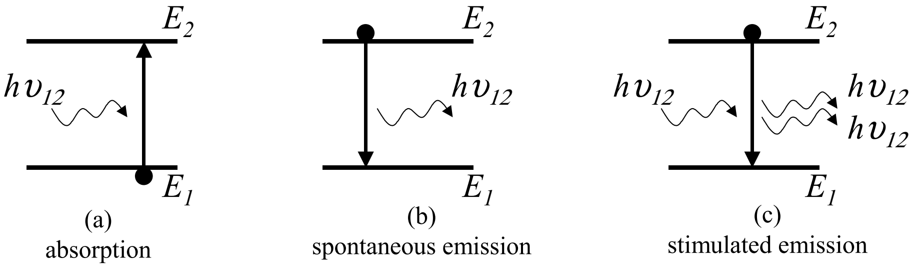
6.3. General laser theory

All the lasers discussed above have several common characteristics. Originally, in order to qualify as a laser, stimulated emission had to be demonstrated. The convention is now that a laser must demonstrate both stimulated emission and positive optical feedback. These concepts will be addressed in this section.

6.3.1. Stimulated emission

Most materials exhibit some kind of optical absorption. Absorption is the process by which incident light is converted to electrical potential energy. For example, an electron can be excited from the valence band to the conduction band by absorbing the energy of a photon. Logically, most materials also exhibit some form of light emission. For example, in an ordinary light source such as a light bulb, the emission of radiation occurs through a process called spontaneous emission. In this process, an electron which had been excited into a higher energy state for "some time" falls down to a lower state by emitting a photon. This "time" is called the radiative recombination lifetime.

However, in the presence of photons, an excited electron can be forced or stimulated to fall down to a lower state much faster than in a spontaneous event. The stimulus is provided by a photon with the proper wavelength. This process is called stimulated emission and produces an additional photon of exactly the same direction of propagation, frequency, and polarization (direction of the electric field in a wave) as the stimulating photon.



(a) absorption

(b) spontaneous emission

(c) stimulated emission

Fig. 6.2. Interaction of photons and electrons in a two energy level system: (a) optical absorption, an electron is excited from the lower level into the upper level by absorbing the energy of a photon of the correct energy; (b) spontaneous emission, the relaxation of the excited electron from the upper level to the lower level by release of energy in the form of a photon; (c) stimulated emission, the relaxation process is triggered by an incident photon and results in an additional photon with the same energy as the incident photon.

Spontaneous and stimulated processes are illustrated in Fig. 6.2. An electron in state E_2 drops spontaneously to E_1 emitting a photon with energy $h\upsilon_{12} = E_2\text{-}E_1$. Assuming that the system is immersed in an intense field of photons, each having energy $h\upsilon_{12} = E_2\text{-}E_1$, the electron is induced to transit from E_2 to E_1, contributing a photon whose wave is *in phase* with the radiation field. If this process continues and other electrons are stimulated to emit photons in the same fashion, a large radiation field due to stimulated emission can build up. This radiation is monochromatic since each photon has an energy $h\upsilon_{12} = E_2 - E_1$ and is coherent because all the photons are released in the same phase.

Let us consider the general conditions necessary for stimulated emission to occur. We assume the populations of energy levels E_1 and E_2 ($E_1 < E_2$) to be n_1 and n_2, respectively. At thermal equilibrium the relative population will be:

Eq. (6.1) $$\frac{n_2}{n_1} = e^{-(E_2-E_1)/k_bT} = e^{-h\nu_{12}/k_bT}$$

At equilibrium most electrons are in the lower energy level, i.e. $n_2 << n_1$. If the atoms exist in a radiation field of photons with energy $h\nu_{12}$ such that the energy density of the radiation field is $\rho(\upsilon_{12})$, i.e. the photon density of states, then stimulated emission can occur along with absorption and spontaneous emission. The rate of stimulated emission is proportional to the number of electrons in the upper level n_2 and to the energy density of the radiation field $\rho(\upsilon_{12})$. It can then be written as $B_{21}n_2\rho(\upsilon_{12})$ where B_{21} is the proportionality coefficient.

The rate at which the electrons make upward transitions from E_1 to E_2 (photon absorption) should also be proportional to $\rho(\upsilon_{12})$ and to the electron population in E_1. This rate is given by $B_{12}n_1\rho(\upsilon_{12})$, where B_{12} is a proportionality factor for upward transitions.

Finally, the rate of spontaneous emission is proportional only to the population of the upper level, $A_{21}n_2$.

The coefficients B_{21}, B_{12}, and A_{21} are called the Einstein coefficients. The ratio of the stimulated to spontaneous emission rates is:

Eq. (6.2) $$\frac{\text{Stimulated emission rate}}{\text{Spontaneous emission rate}} = \frac{B_{21}n_2\rho(\upsilon_{12})}{A_{21}n_2} = \frac{B_{21}}{A_{21}}\rho(\upsilon_{12})$$

which is generally very small, so the contribution of stimulated emission is negligible. From Eq. (6.2) it follows that in order to enhance the stimulated emission over spontaneous emission one has to provide large photon field energy density $\rho(\upsilon_{12})$. In a laser, this is encouraged by providing a resonant optical cavity in which the photon density can build up to a large value through multiple internal reflections.

Under thermal equilibrium the total rates of upward and downward transitions are equal:

Eq. (6.3) $$B_{12}n_1\rho(\upsilon_{12}) = A_{21}n_2 + B_{21}n_2\rho(\upsilon_{12})$$

Eq. (6.3) means that in equilibrium the ratio of the stimulated emission and upward transition rates is given by:

Eq. (6.4) $$\frac{\text{Stimulated emission rate}}{\text{Absorption rate}} = \frac{B_{21}}{B_{12}}\frac{n_2}{n_1} < 1$$

So, stimulated emission may dominate over absorption only when, (if $B_{12} = B_{21}$) $n_2 > n_1$, i.e. in a non-equilibrium state. This state is called population inversion. In summary, Eq. (6.2) and Eq. (6.3) indicate that stimulated emission can dominate if two requirements are met: (i) there is an optical resonant cavity to encourage the photon field to build up, (ii) there is population inversion.

The first requirement implies multiple passes of the light through the amplifying medium and the second requirement is a necessary condition for the medium to amplify the input light (Fig. 6.2(c)). Fig. 6.3 illustrates these two conditions for the generation of high intensity, coherent light.

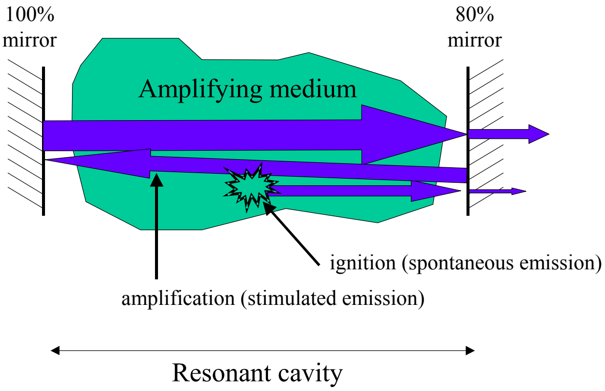

100% mirror
80% mirror
Amplifying medium
ignition (spontaneous emission)
amplification (stimulated emission)
Resonant cavity

Fig. 6.3. The elements necessary to a laser system: a resonant cavity which ensures the buildup of a photon field, an amplifying medium in which the population inversion has been established, and an external pump which supplies the energy necessary to the system for amplification.

6.3.2. Resonant cavity

As shown in Fig. 6.4, light propagating inside the amplifying medium and perpendicularly to the mirrors (on-axis) with a particular frequency can be reflected back and forth within the resonant cavity in a reinforcing or coherent manner if an integral number of half-wavelengths fit between the end mirrors. Thus the length of the cavity for stimulated emission must be:

Eq. (6.5) $$L = \frac{m\lambda}{2}$$

where m is an integer. In this equation, λ is the photon wavelength within the laser material. It should be noted that this wavelength is related to the vacuum wavelength λ_0 through the relation:

Eq. (6.6) $$\lambda_0 = \lambda\bar{n}$$

where $\bar{n}$ is the refractive index of the resonant cavity material. The allowed wavelengths within the optical cavity are referred to as longitudinal optical modes.

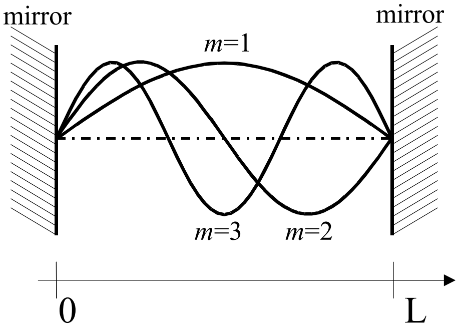


Fig. 6.4. Longitudinal optical modes in an optical resonant cavity delimited by two parallel mirrors. The length of the cavity is equal to an integral number of half-wavelengths of the light being amplified. This configuration leads to the constructive interference or reinforcing reflection of the light by the cavity, and thus amplification.

The amplifying medium is characterized by a definite wavelength region in which stimulated emission can occur. This is referred to as the material gain curve and is shown in Fig. 6.5(a). In a given resonant cavity including this amplifying medium, only the longitudinal modes will experience amplification, which leads to a characteristic laser output as shown in Fig. 6.5(c).

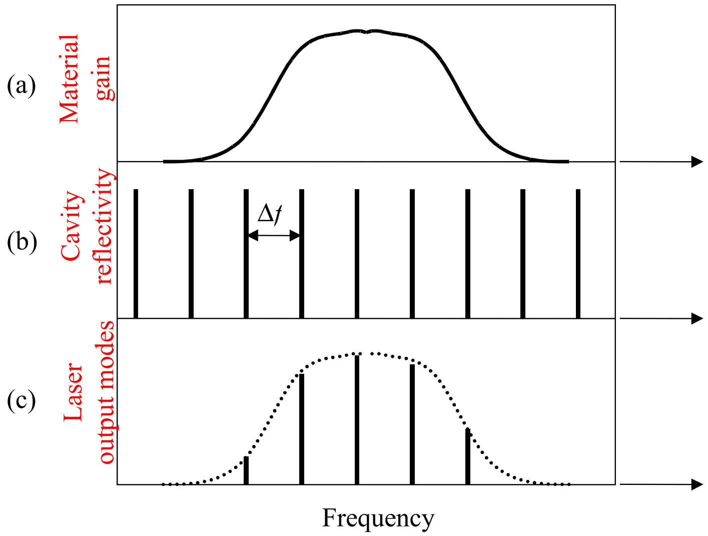


Fig. 6.5. Illustration of the effect of the cavity longitudinal modes on the laser output spectrum: (a) gain curve of the amplifying medium, (b) the wavelengths of the allowed longitudinal modes inside a given resonant cavity, and (c) the laser output modes.

Light output from a laser is also characterized by off-axis transverse modes. The transverse modes refer to the spatial intensity distribution at the exit mirrors. The origin of these modes is similar to the longitudinal modes, but is related to the spatial interference of light in the cavity. Fundamental modes are typically most intense on the axis of the cavity. Higher order modes exhibit multiple intensity peaks of increasing spatial frequency. Semiconductor lasers in particular demonstrate very specific transverse modes.

6.3.3. Waveguides

In order to achieve low power consumption and high efficiency, modern semiconductor lasers include thin layers (<1 μm) deposited by epitaxial techniques (Chapter 1) such that the electrons and holes are confined into a narrow region. In addition, when different types of materials with different refractive indices are used, these layers can also confine light, i.e. constitute a waveguide which is a medium in which a wave can propagate in one direction and is confined in others.

Light is an electromagnetic wave, composed of oscillatory electric and magnetic fields perpendicular to each other and the direction of propagation. The propagation of light in a medium can be described by Maxwell's equations which give the relation between the electric and magnetic fields in the wave. In this part, the propagation of the light inside a dielectric waveguide will be discussed.

Maxwell's equations can be simplified in a dielectric since there is a negligible electrical current inside such a material. In the single frequency mode approximation and source free region, they are:

Eq. (6.7)
$$\begin{cases} \vec{\nabla}\times\vec{E} = i\omega\vec{B} \\ \vec{\nabla}\times\vec{H} = -i\omega\vec{D} \\ \vec{\nabla}\cdot\vec{B} = 0 \\ \vec{\nabla}\cdot\vec{D} = 0 \end{cases}$$

where $\vec{E}$ is the electric field, $\vec{B}$ is the magnetic induction or flux density, $\vec{H}$ is the magnetic field strength, $\vec{D}$ is the electric displacement, and ω is the angular frequency of the electromagnetic wave (i.e. light). In an isotropic material, the displacement $\vec{D}$ and electric field strength $\vec{E}$ are related through the absolute permittivity ε of the material (in this Chapter the permeability is always the absolute permeability):

Eq. (6.8)
$$\vec{D} = \varepsilon\vec{E}$$

Similarly, the magnetic field strength $\vec{H}$ and the magnetic flux density $\vec{B}$ are related through the permeability μ of the material:

Eq. (6.9)
$$\vec{B} = \mu\vec{H}$$

Using Eq. (6.9) in the first relation in Eq. (6.7), we have:

Eq. (6.10)
$$\vec{\nabla}\times\vec{E} = i\omega\left(\mu\vec{H}\right)$$

After applying $\vec{\nabla}\times$ to both sides of Eq. (6.10), we get:

Eq. (6.11)
$$\vec{\nabla}\times\left(\vec{\nabla}\times\vec{E}\right) = i\omega\mu\left(\vec{\nabla}\times\vec{H}\right)$$

On the other hand, we have the mathematical relation:

Eq. (6.12)
$$\vec{\nabla}\times\left(\vec{\nabla}\times\vec{E}\right) = \vec{\nabla}\times\left(\vec{\nabla}\cdot\vec{E}\right) - \nabla^2\vec{E}$$

where ∇^2 represents the Laplacian operator and is such that:

Eq. (6.13) $$\nabla^2\vec{E}=\frac{\partial^2\vec{E}}{\partial x^2}+\frac{\partial^2\vec{E}}{\partial y^2}+\frac{\partial^2\vec{E}}{\partial z^2}$$

In addition, using Eq. (6.8) in the last relation in Eq. (6.7), we get:

Eq. (6.14) $$\vec{\nabla}\cdot\vec{D}=\vec{\nabla}\cdot\left(\varepsilon\vec{E}\right)=\varepsilon\vec{\nabla}\cdot\vec{E}=0$$

which means, because the permittivity ε cannot be zero, that:

Eq. (6.15) $$\vec{\nabla}\cdot\vec{E}=0$$

Combining Eq. (6.11), Eq. (6.12) and Eq. (6.15) we find that:

Eq. (6.16) $$-\nabla^2\vec{E}=i\omega\mu\left(\vec{\nabla}\times\vec{H}\right)$$

Now, by inserting the second relation of Eq. (6.7) into this last equation, we get:

Eq. (6.17) $$-\nabla^2\vec{E}=i\omega\mu\left(-i\omega\vec{D}\right)=\mu\omega^2\vec{D}$$

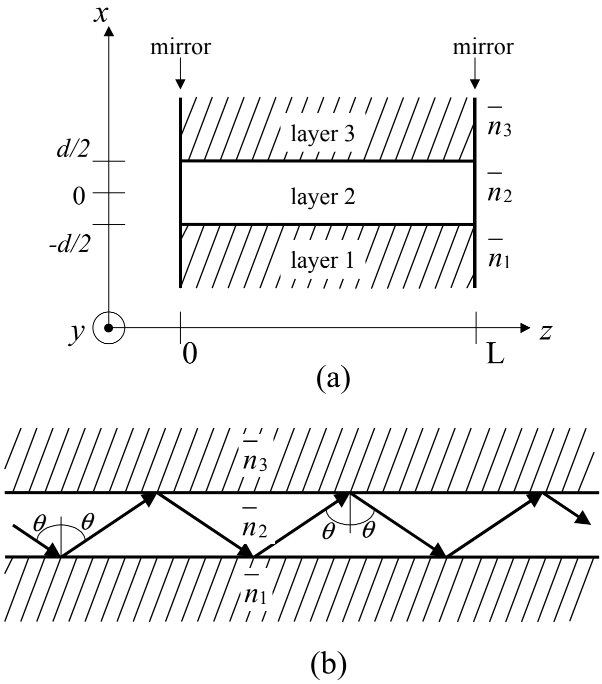
Using Eq. (6.8), this becomes:

Eq. (6.18) $$\nabla^2\vec{E}+\mu\varepsilon\omega^2\vec{E}=\vec{0}$$

This equation is sometimes called the wave equation and governs the behavior of the electric field strength component $\vec{E}$ of an electromagnetic wave (i.e. light) in a medium. Knowing $\vec{E}$, one can use the first relation in Eq. (6.7) to calculate $\vec{B}$, the magnetic component of the propagating wave.

Solving the wave equation in any such non-isotropic structure also requires the knowledge of the boundary conditions. These are in part determined by the geometry of the waveguide. The wave equation is of the second order in $\vec{E}$ and, for a symmetric waveguide, yields even and odd parity solutions, which are called even and odd modes.

The simplest waveguide is a slab waveguide and consists of a high index core layer sandwiched by two other parallel layers, called cladding layers, with different refractive indices, as shown in Fig. 6.6. As a result, the propagation of the electromagnetic wave depends on whether the electric field is parallel to the layers or perpendicular to them. The former case is called TE polarization while the latter is called TM polarization. The notation ‘transverse electric’ (TE) and ‘transverse magnetic’ (TM) refer to the field direction relative to the plane of reflection as shown in Fig. 6.6(b).



(a)

(b)

Fig. 6.6. (a) Representation of a three-layer dielectric waveguide, with three different refractive indices. (b) Ray trajectories of the guided wave, when the refractive index of the center layer is larger than those of the surrounding layers. The ray of light can experience total internal reflection at the interfaces between the dielectric materials, confining light to the core material.

To model the wave propagation in such a structure, one needs to consider two general solutions to the wave equation Eq. (6.18). The fundamental mode of most semiconductor lasers is TE. A TE mode must satisfy the wave equation and takes on the general form in the waveguide core:

Eq. (6.19) $$E_y(x,z)=A\exp(ik_z z)\cos(k_x x+\varphi)$$

where A is a normalization constant, and:

Eq. (6.20) $k_x{}^2 + k_z{}^2 = \omega^2 \mu\varepsilon$

where k_z is the propagation wavevector in the waveguide, ε is the permittivity and μ the permeability of the waveguide. In the slab waveguide, some values of k_z cause k_x to become imaginary. This leads to a decay of the mode in the waveguide cladding, and is called the evanescent solution:

Eq. (6.21) $E_y(x,z) = \exp(ik_z z)\left[B\exp(-\alpha x) + C\exp(\alpha x)\right]$

where B and C are normalization constants, and:

Eq. (6.22) $-\alpha^2 + k_z^2 = \omega^2\mu\varepsilon$

and:

Eq. (6.23) $\alpha = ik_x$

Let us assume that the waveguide is limited in the x-direction to within the region ($-d/2 < x < d/2$), that it extends to infinity in the y-direction and that the waves propagate in the z-direction, as shown in Fig. 6.6. By solving the wave equation, one can find that the electric field in the even TE-polarization mode is along the y-direction and is given by:

Eq. (6.24)

$$E_y(x,z) = e^{ik_z z}\begin{cases} A_1 \exp\left(-\alpha_1\left(x - \dfrac{d}{2}\right)\right) & for\ x > \dfrac{d}{2} \\ A_2 \cos(k_{2x}x + \phi) & for\ -\dfrac{d}{2} < x < \dfrac{d}{2} \\ A_3 \exp\left(\alpha_3\left(x + \dfrac{d}{2}\right)\right) & for\ x < -\dfrac{d}{2} \end{cases}$$

where (α_1, k_{2x}, α_3) are the components of the wavevector in the x-direction defined by:

Eq. (6.25) $$\begin{cases} k_{2x}{}^2 + k_z^2 = \omega^2 \mu_2 \varepsilon_2 \\ -\alpha_1{}^2 + k_z^2 = \omega^2 \mu_1 \varepsilon_1 \\ -\alpha_3{}^2 + k_z^2 = \omega^2 \mu_3 \varepsilon_3 \end{cases}$$

and k_z is the propagation wavevector for the confined mode.

The magnetic field strength can be determined from the electric field strength using Eq. (6.10), given that $\vec{E} = E_y \vec{y}$:

Eq. (6.26) $$\vec{H} = \frac{1}{i\omega\mu_l}\left(-ik_z E_y \vec{x} + \frac{\partial E_y}{\partial x}\vec{z}\right)$$

where μ_l is the permeability of the material in layer l. The boundary conditions for the electric and magnetic field strengths in a waveguide rely on the continuity of their tangential components, i.e. E_y and H_z. Applying these boundary conditions at $x = \pm d/2$ yields a transcendental equation, similar to finding the bound states for an electron in a finite potential well. By further assuming the permeabilities are such that $\mu_1 = \mu_2 = \mu_3$, which is the case for most III-V semiconductors, this transcendental equation has the form:

Eq. (6.27) $$\tan(k_{2x} d) = \frac{(\alpha_1 + \alpha_3) k_{2x}}{k_{2x}{}^2 - \alpha_1 \alpha_3}$$

Using Eq. (6.25), this can be solved graphically as a function of k_{2x}, as shown in Fig. 6.7.

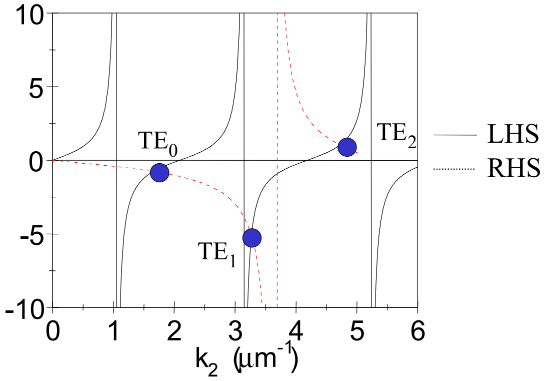


Fig. 6.7. Plot of the left-hand side (LHS) and right-hand side (RHS) of Eq. (6.27). The intersection points represent allowed optical modes referenced as TE$_0$ through TE$_2$.

The complete expression of the electric field strength can be obtained after applying the normalization condition for the power propagated by this mode:

Eq. (6.28) $$P = -\frac{1}{2}\mathrm{Re}\int_{-\infty}^{\infty}\left(E_y H_x^{\ *}\right)dx = 1$$

where $H_x^{\ *}$ is the complex conjugate of the quantity H_x. Note the negative sign simply means that for positive Ey and Hx the power propagates to the left. This gives:

Eq. (6.29)
$$E_y(x,z) = Ae^{ik_z z}\begin{cases}\cos\left(k_{2x}\frac{d}{2}+\phi\right)\exp\left(-\alpha_1\left(x-\frac{d}{2}\right)\right) & for\, x > \frac{d}{2} \\ \cos(k_{2x}x+\phi) & for\, -\frac{d}{2} < x < \frac{d}{2} \\ \cos\left(-k_{2x}\frac{d}{2}+\phi\right)\exp\left(\alpha_3\left(x+\frac{d}{2}\right)\right) & for\, x < -\frac{d}{2}\end{cases}$$

where:

Eq. (6.30) $$A = \sqrt{\frac{4\omega\mu}{k_z\left(d + \frac{1}{\alpha_1} + \frac{1}{\alpha_3}\right)}}$$

and:

Eq. (6.31) $$\phi = \cot^{-1}\left(\frac{k_{2x}}{\alpha_1}\right) - k_{2x}\frac{d}{2}$$

Each of these TE modes solutions of the above equations will be indexed with an integer p and has its own wavenumber k_z, varying from that in the cladding layer $k_z = \frac{\omega\bar{n}_1}{c}$ at low frequency to that in the core $k_z = \frac{\omega\bar{n}_2}{c}$ at high frequency.

When $\bar{n}_1 = \bar{n}_3 = \bar{n}_{clad}$ and $\bar{n}_2 = \bar{n}_{core}$, the slab waveguide is said to be symmetric. In this case, the allowed modes take on an even or odd parity and look similar to Fig. 6.8. All the TE_p modes, for $p > 0$, exhibit a cutoff frequency, ω_p, such that a wave with a lower frequency cannot propagate through the waveguide in that mode. For the symmetric waveguide, the predictable cutoff condition is:

Eq. (6.32) $$\frac{\omega}{c}\frac{d}{2}\sqrt{\bar{n}_{core}^2 - \bar{n}_{clad}^2} = p\frac{\pi}{2} \qquad p = 0,1,2,...$$

which indicates that the TE_0 mode always has a solution, while higher order modes may not.

The equations for TM modes are derived in a similar fashion to the TE. This is done by using the duality principle, which consists of replacing $\vec{E}$ and $\vec{H}$ with $\vec{H}$ and $-\vec{E}$, respectively, and swapping μ with ε in all the formulas. It should be realized that, in this case, E_x will be discontinuous due the change of permittivity at material interfaces.

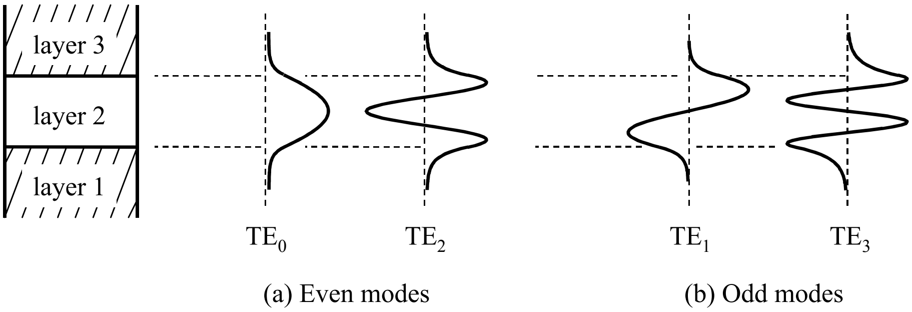


Fig. 6.8. Electric field profiles in TE modes in a symmetric waveguide: (a) even modes, (b) odd modes.

In practice, the waveguide in a semiconductor laser is formed by an active layer or active region surrounded by two cladding layers of different material. These layers typically have a smaller refractive index than the active layer, which confines the light within the active layer and guides it.

The extent of the confinement is mathematically described by the optical confinement factor, Γ, which in turn influences the threshold current and other laser characteristics. In terms of the above solution for TE modes, the confinement factor can be expressed as:

Eq. (6.33)
$$\Gamma = \frac{-\frac{1}{2}\operatorname{Re}\int_{-d/2}^{d/2}\left(E_y H_x^{*}\right)dx}{-\frac{1}{2}\operatorname{Re}\int_{-\infty}^{\infty}\left(E_y H_x^{*}\right)dx}$$

where E_y and H_x are the components of the electric and magnetic field strengths in the y- and x-directions, respectively. This represents the percentage of the optical mode power that is confined within the waveguide core.

A waveguide with a high confinement factor makes efficient use of emitted light and tends to have a low threshold gain/current. For a symmetric waveguide, a thick core and/or high refractive index difference between layers leads to a high confinement factor. This effect is shown graphically in Fig. 6.9.

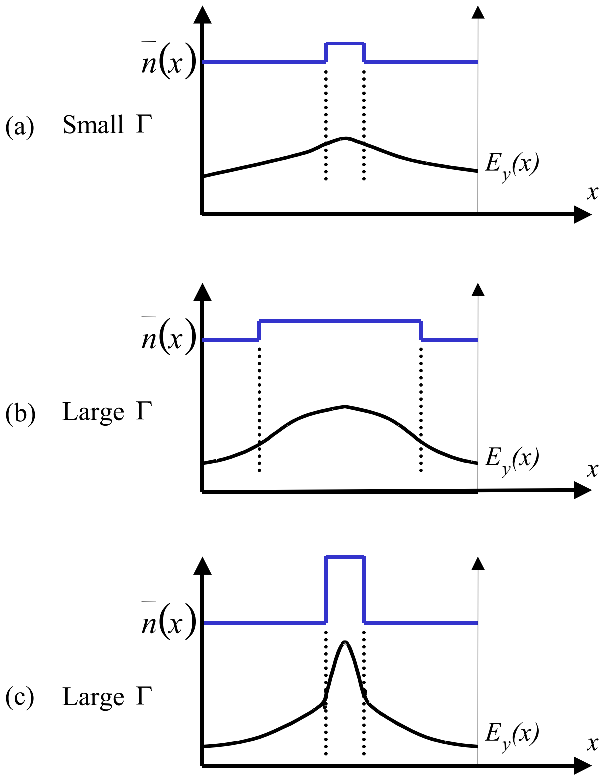


Fig. 6.9. TE$_0$ mode and confinement factor dependence on the refractive index difference $\Delta\bar{n}$ and core thickness d. (a) small d, small $\Delta\bar{n}$, small Γ; (b) large d, small $\Delta\bar{n}$, large Γ; (c) small d, large $\Delta\bar{n}$, large Γ.

6.3.4. Laser propagation and beam divergence

Regardless of the optical confinement, laser light exiting a semiconductor laser ($\bar{n} \approx 3.4$) into air ($\bar{n} = 1$) diverges in the x-direction due to diffraction. This happens for all lasers, but due to their small emitting aperture, the effect is much more pronounced in semiconductor lasers.

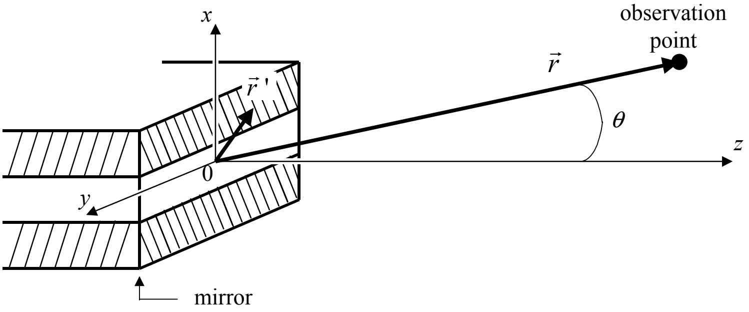


Fig. 6.10. Coordinates for calculating the far-field pattern at a distance r.

Let us consider the laser geometry shown in Fig. 6.10. At some distant observation point from the mirror ($z = 0$), such that $r >> r'$, the far-field approximation holds, and is given as:

Eq. (6.34) $$\left|\vec{r} - \vec{r}'\right| \approx \left|\vec{r}\right| - \vec{r}' \cdot \frac{\vec{r}}{\left|\vec{r}\right|}$$

Following a derivation from Chuang [1995] based on the propagating radiation field, and the definition of the radiated power $I = \dfrac{1}{2\sqrt{\dfrac{\mu_0}{\varepsilon_0}}}\left|E(\vec{r})\right|^2$ the angular dependence of the power distribution for a TE mode is given by the following transform:

Eq. (6.35) $$I(\theta)_\perp \propto \cos^2\theta \left|\int_{-\infty}^{\infty} e^{-ikx'\sin\theta} E_y(x', z = 0)dx'\right|^2$$

This intensity distribution is called the far-field pattern. The integral in Eq. (6.35) is effectively a spatial Fourier transform, as a small aperture translates to a large divergence.

In a symmetric slab waveguide characterized by a core thickness of d, a Gaussian function can be used to approximate the TE_0 transverse mode. This leads to an effective spot size, s_e, given by:

Eq. (6.36)
$$s_e = \begin{cases} \dfrac{d}{2}\left[1+\dfrac{2}{\alpha d}\ln\left(\sqrt{e}\cos\left(\dfrac{k_x d}{2}\right)\right)\right] & \mathit{for}\ \cos\left(\dfrac{k_x d}{2}\right) > \dfrac{1}{\sqrt{e}} \\ d\left[\dfrac{1}{\alpha d}\cos^{-1}\left(\dfrac{1}{\sqrt{e}}\right)\right] & \mathit{for}\ \cos\left(\dfrac{k_x d}{2}\right) < \dfrac{1}{\sqrt{e}} \end{cases}$$

This function is easier to transform, and leads to a simpler solution for the divergence angle, $\theta_\perp$, given by:

Eq. (6.37)
$$\theta_\perp \approx \frac{2}{ks_e} = \frac{\lambda_0}{\pi s_e}$$

In other words, the divergence angle is inversely proportional to the spot size. This is very important with regard to application. In general, the higher the divergence, the harder it is to collect all the light.

Unlike the y-direction of a slab waveguide, the y-direction of a real laser diode is not truly uniform. Various gain- and index-guided designs are used to control the transverse mode output in this direction as well, which makes

Eq. (6.38)
$$E_y(x,z) \rightarrow E_y(x,y,z)$$

As a consequence, even in the transverse direction of the waveguide, there is divergence, as shown in Fig. 6.11.

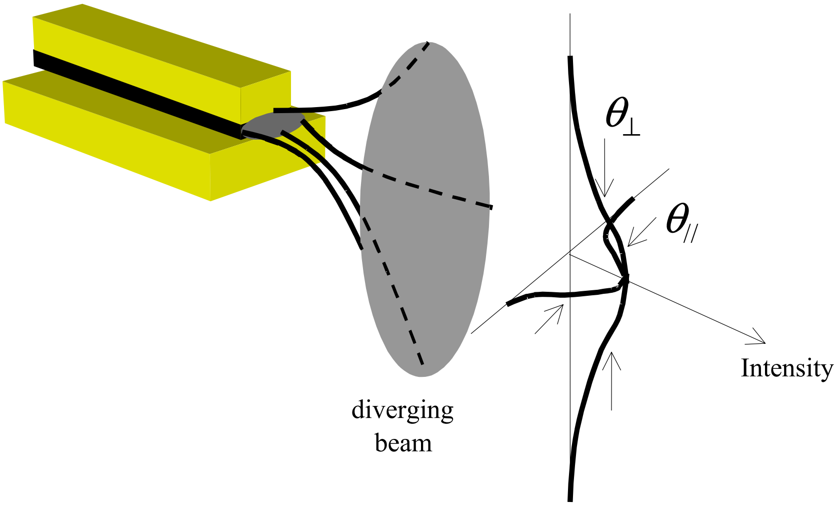


Fig. 6.11. Divergence characteristic of a non-infinite semiconductor laser waveguide.

The resulting power intensity pattern in the two directions can be given as:

Eq. (6.39)
$$\begin{cases} I\left(\theta_{\perp}\right) \propto \cos^2\theta \left| \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} e^{-ikx'\sin\theta} E_y(x',y',z=0)\,dx'dy' \right|^2 \\ I\left(\theta_{\parallel}\right) \propto \left| \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} e^{-iky'\sin\theta} E_y(x',y',z=0)\,dx'dy' \right|^2 \end{cases}$$

6.3.5. Waveguide design considerations

It should be noted that higher order transverse modes will all have different divergence characteristics. When multiple modes exist, interference effects can produce non-Gaussian, poor quality, far-field patterns. This condition makes it extremely difficult to focus the laser output to a tight spot, which decreases the usable output.

The waveguide needs to be designed for a high confinement factor, low divergence, and good beam quality. Care must be taken to suppress higher order transverse modes, which appear more readily when there is a large index difference or a thick waveguide core. For the most usable output, a laser of given frequency ω, needs a waveguide such that $\omega < \omega_1$, where ω_1 is the cutoff frequency for the TE_1 mode.

For semiconductor lasers in general, even when higher modes do exist, it is the ellipticity of the semiconductor laser output, as shown in Fig. 6.11~~,~~ that makes collection and focusing of light difficult. This is especially true for laser arrays in which there may be a very large aspect ratio to overcome. Cylindrical lenses are needed to first circularize the beam before the output can be utilized. A laser designed with a circular output can take advantage of simpler, and in general, high quality optics.

6.4. Ruby laser

The first working laser was built in 1960 by Maiman, using a ruby crystal as the amplifying or active medium. Ruby belongs to the family of gems consisting of sapphire or alumina ($A1_2O_3$) with various types of impurities. For example, pink ruby contains about 0.05% Cr atoms. Similarly, $A1_2O_3$ doped with Ti, Fe, or Mn results in variously colored sapphire. Most of these materials can be grown as single crystals.

Ruby crystals are available in rods several inches long, convenient for forming an optical cavity (Fig. 6.12). The crystal is cut and polished so that the ends are flat and parallel, with the end planes perpendicular to the axis of the rod. These ends are coated with a highly reflective material, such as Al or Ag, producing a resonant cavity in which light intensity can build up through multiple reflections. One of the end mirrors is constructed to be partially transparent so that a fraction of the light will “leak out” of the resonant system. This transmitted light is the output of the laser. Of course, in designing such a laser one must choose the amount of transmission to be a small perturbation on the resonant system. The gain in photons per pass between the end plates must be larger than the transmission at the ends, as well as any other losses due to light scattering and absorption. The arrangement of parallel plates providing multiple internal reflections is similar to that used in the Fabry-Perot interferometer; thus the silvered ends of the laser cavity are often referred to as Fabry Perot faces.

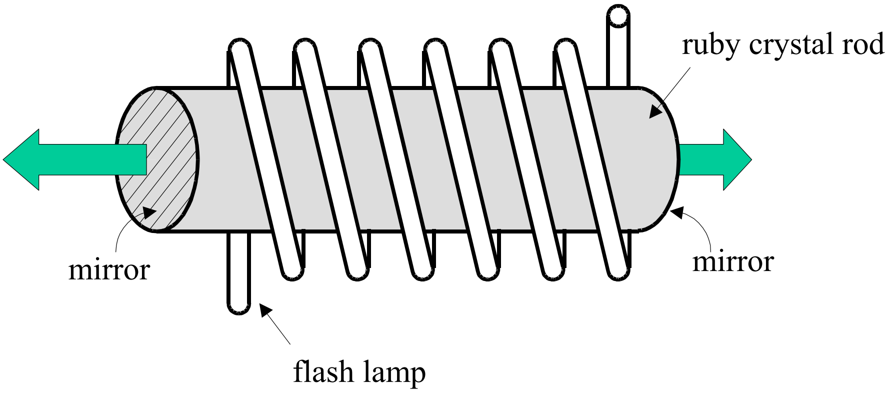


Fig. 6.12. Schematic diagram of a ruby laser. A ruby crystal rod is cut and polished so that its ends form mirrors to create the resonant cavity. A flash lamp supplies the necessary energy in the form of photons to pump the rod.

In the case of ruby, chromium (Cr) atoms in the crystal have their energy levels as shown in Fig. 6.13, where only the energy levels that are important for stimulated emission are depicted.

This is basically a three-level system. Absorption occurs in the green part of the spectrum, exciting electrons from the ground state E_1 to the band of levels designated E_3 in the figure. Then electrons decay rapidly to the level E_2. This transition is non-radiative. The level E_2 is very important for the stimulated emission process since electrons in this level have a mean lifetime of about 5 ns before they fall to the ground state. Because this lifetime is relatively long, E_2 is called a metastable state. If electrons are excited from E_1 to E_3 at a rate faster than the radiative rate from E_2 back to E_1, the population of the metastable state E_2 becomes larger than that of the ground state E_1 (we assume that electrons fall from E_3 to E_2 in a negligibly short time).

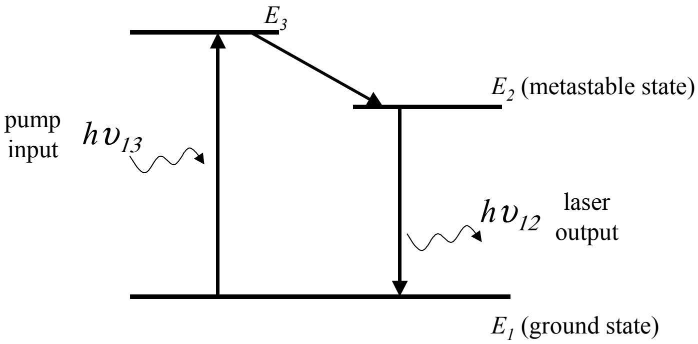


Fig. 6.13. Energy levels for chromium ions in ruby. The three level system includes a ground level at E_1, an excited level at E_3, and a metastable state at E_2 where the excited electrons relax rapidly to. The mean lifetime of the metastable state is long enough to ensure that population inversion can be achieved between the levels E_1 and E_2.

In the experiment done by Maiman in 1960, population inversion is obtained by optical pumping of the ruby rod with a flash lamp such as the one shown in Fig. 6.12. A common type of flash lamp is a glass tube wrapped around the ruby rod and filled with xenon gas. A capacitor can be discharged through the xenon-filled tube, creating a pulse of very intense light over a broad spectral range. If the light pulse from the flash tube is several milliseconds in duration, we might expect an output from the ruby laser over a large fraction of that time. However, the laser does not operate continuously during the light pulse but instead emits a series of very short spikes (Fig. 6.14).

When the flash lamp intensity becomes large enough to create population inversion (the threshold pumping level*),* stimulated emission from the metastable level to the ground level occurs, with a resulting laser emission. Once the stimulated emission begins, the metastable level is depopulated very quickly. Thus the laser output consists of an intense spike lasting from a few nanoseconds to microseconds. After the stimulated emission spike, population inversion builds up again and a second spike results. This process continues as long as the flash lamp intensity is above the threshold pumping level.

In this situation, one can easily understand that the metastable level never receives a highly inverted population of electrons. Whenever the population of E_2 reaches the minimum required for stimulated emission, these electrons are depleted quickly in one of the laser emission spikes.

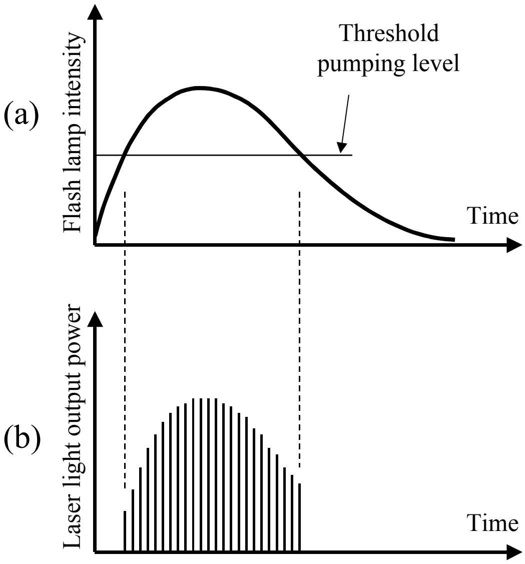


Fig. 6.14. Laser spikes in the output of a ruby laser: (a) typical variation of the intensity of the flash lamp with time: the intensity is above the threshold pumping level only during a certain period of time, not during the entire duration when the lamp is powered. (b) Laser spikes occurring while the flash intensity is above the threshold pumping level. As soon as population inversion is achieved between the levels E_1 and E_2, the laser emits a pulse of light. This process results in the series of laser intensity spikes.

To prevent this, we must somehow keep the coherent photon field in the ruby rod from building up (and thus prevent stimulated emission) until after a larger population inversion is obtained. This can be accomplished if we temporarily interrupt the resonant character of the optical cavity.

This process is called Q-switching, where Q is the quality factor of the resonant structure. A straightforward method for doing this is illustrated in Fig. 6.15. The front face of the ruby rod is silvered to be partially reflecting, but the back face is left un-silvered. The back reflector of the optical cavity is provided by an external mirror, which can be rotated at high speeds. When the mirror plane is aligned exactly perpendicular to the laser axis, a resonant structure exists; but as the mirror rotates away from this position, there is no buildup of photons through multiple reflections, and no laser action can occur. Thus during a flash from the xenon lamp, a very large inverted population builds up while the mirror rotates off-axis. When the mirror finally returns to the position at which light reflects back into the rod, stimulated emission can occur, and the large population of the metastable level is given up in one intense laser pulse. This structure is called a giant pulse laser or a Q-switched laser. By saving the electron population for a single pulse, a large amount of energy can be given up in a very short time.

For example, if the total energy in the pulse is 1 Joule and the pulse width is 100 ns (10^{-7} s), the peak pulse power is 10^7 J·s^{-1} = 10 MW.

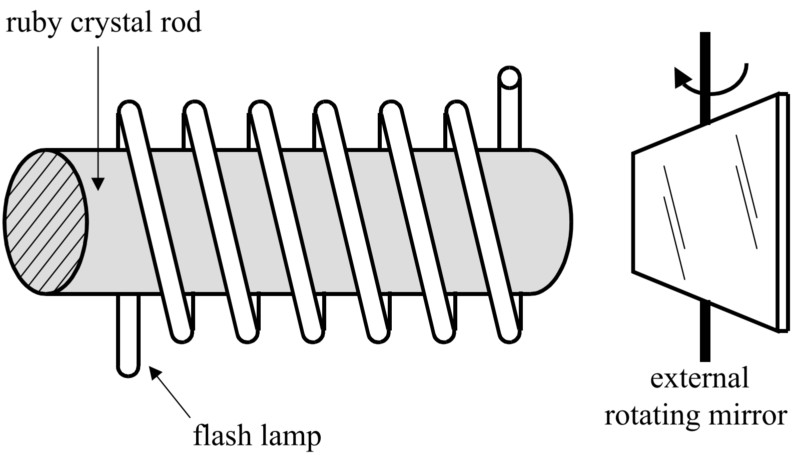


Fig. 6.15. Schematic diagram of a Q-switched ruby laser in which one face of the resonant cavity is an external rotating mirror. The purpose of the rotating mirror is to prevent stimulated emission by interrupting the resonant nature of the laser cavity, thus preventing the photon field from building up. This allows a larger population inversion to be achieved, and consequently, a higher peak intensity laser light emission.

6.5. Semiconductor lasers

In a semiconductor laser, population inversion mechanism is realized through a very unique method: by injecting electrical current directly into a *p-n* junction. This method of achieving population inversion is very efficient when compared to the process in ruby lasers or gas lasers. The semiconductor laser itself is also very compact (a typical size of the active laser part is only $100\,\mu\text{m} \times 1000\,\mu\text{m} \times 100\,\mu\text{m}$ = one part in a hundred thousand cubic centimeters!).

Moreover, semiconductor lasers can be easily integrated with other types of semiconductor devices such as transistors or even large-scale integrated circuits, and the laser output can be easily modulated by controlling the junction current. It is no surprise that semiconductor lasers are now widely used for high speed optical processing and optical communication.

Another great advantage for these lasers is an inherent optical cavity. Most popular semiconductors (III-V, II-VI) have natural cleavage planes, which are the crystallographic planes along which the atomic bonds are weakest and therefore most easily broken. For zinc-blende crystals, cleavage parallel to (110) and $(1\bar{1}0)$ planes can produce atomically flat mirrors for use in a Fabry-Perot optical cavity. The reflectivity of the mirrors is limited by the refractive index of the semiconductor, and is given by:

Eq. (6.40)
$$R = \frac{\left(\bar{n}-1\right)^2}{\left(\bar{n}+1\right)^2}$$

where $\bar{n}$ is the refractive index of the semiconductor. A typical semiconductor has a refractive index of 3.4 in the mid-infrared region, which gives a natural mirror reflectivity of 29.8%.

6.5.1. Population inversion

If a *p-n* junction is formed between degenerate materials, the bands under forward bias appear as shown in Fig. 6.16. If the injected current is large enough, electrons and holes are injected into and travel across the transition region in considerable concentrations. A large concentration of electrons is then present in the conduction band, while a large concentration of holes is present in the valence band, which satisfies the condition for population inversion.

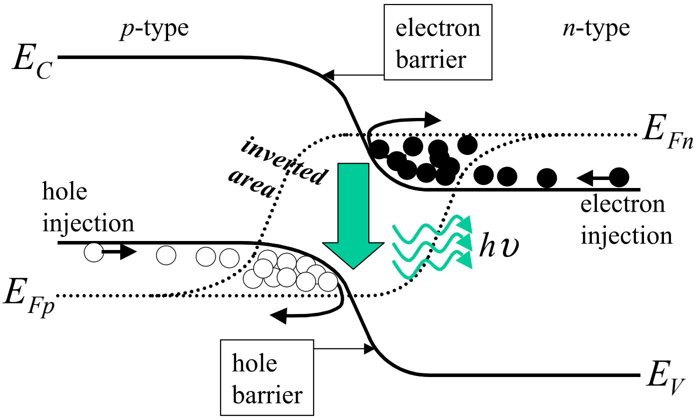


Fig. 6.16. Band diagram of a p-n junction laser under forward bias. The electrons and holes injected into the space charge region recombine radiatively to emit photons with an energy close to the bandgap of the semiconductor in the inversion region. When enough electrons and holes are injected, population inversion can be achieved, making laser emission possible.

Unlike the case of the three-level system discussed earlier (the ruby laser is essentially a three-level system), the condition for population inversion in semiconductors is more complicated. Both electrons and holes experience strong intraband scattering. The rapid intraband scattering and separate injection of electrons and holes allows for thermal equilibrium in each band separately. This situation is often called quasi-equilibrium and the particle distributions are described by Fermi distribution functions using quasi-Fermi energies for electrons and holes E_{F_n} and E_{F_p}.

Let us consider the population inversion in a semiconductor in more detail. We saw in Chapter 4 that, when an external bias was applied, minority carriers are injected on either side of the *p-n* junction and the quasi-Fermi levels go deep into each band apart from their equilibrium position somewhere in the bandgap. The absorption coefficient depends on the quasi-Fermi energies E_{F_n} and E_{F_p}. It changes sign when $E_{F_n} - E_{F_p} = h\upsilon$, for a given photon energy $h\upsilon$.

As the quasi-Fermi levels move apart from each other, this leads to a negative absorption coefficient which means the medium amplifies the light of frequency υ. The condition $E_{F_n} - E_{F_p} = h\upsilon$ is known as the Bernard-Durafforg condition, or transparency point, because at this point the absorption coefficient is zero. Reaching the transparency point, or population inversion, is a necessary condition for lasing. When $E_{F_n} - E_{F_p} > h\upsilon$, light of frequency υ is subject to amplification and is characterized by a gain, which is the opposite of the absorption coefficient.

As a result of this condition, the frequency of the light emitted by semiconductor lasers is larger than $\frac{E_g}{h}$. This in turn means that for lasing to occur in semiconductor lasers, it is necessary to apply a voltage $V = \frac{E_{F_n} - E_{F_p}}{q}$ at least higher than $\frac{E_g}{q}$.

6.5.2. Threshold condition and output power

The photon wavelengths which participate in stimulated emission are determined by the length of the resonant cavity as described in Eq. (6.5). Fig. 6.17 illustrates a typical plot of the light emission intensity versus photon energy for a semiconductor laser.

At low current levels, a spontaneous emission spectrum is observed, as shown in Fig. 6.17(a). As the current is increased to the threshold value, stimulated emission occurs at light frequencies corresponding to the cavity modes as shown in Fig. 6.17(b). Finally, at a higher current level, a most preferred mode or set of modes will dominate the spectral output, as shown in Fig. 6.17(c).

This very intense emission represents the main laser output of the device, where the output light will be composed of almost monochromatic radiation superimposed on a relatively weak radiation background, due primarily to spontaneous emission.

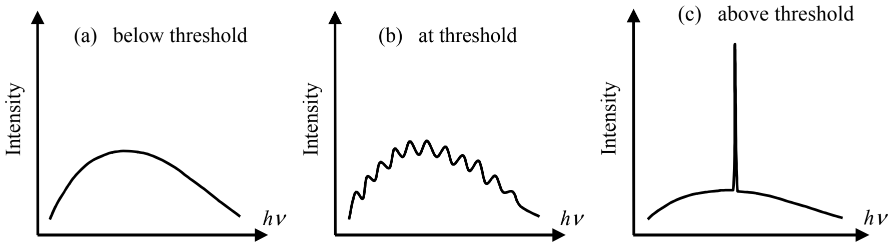


Fig. 6.17. Emission spectrum plotted as light intensity versus the energy of photons for a semiconductor laser: (a) below the threshold, an incoherent emission occurs with many photons emitted at several values of energy; (b) at threshold, laser modes appear which are determined by the dimensions of the resonant cavity; (c) above threshold, one dominant laser mode remains.

When the transparency condition is satisfied, the region becomes active, which means it can amplify light. The peak value g of the gain as function of frequency plays a major role in laser action. Typically, the peak gain is a linear function of carrier density:

Eq. (6.41) $$g(n) = a(n - n_0)$$

where n_0 is the transparency density and a is called the differential gain.

Although gain leads to light amplification, the optical losses, such as absorption outside the active region, α_i, and mirror loss, α_m, prevent the domination of the stimulated emission. So, in real lasers, the threshold current I_{th} has to provide not only the transparency condition but also has to compensate the optical losses in the laser cavity. This means that the real threshold current is larger than that which simply maintains the population inversion. In other words, the threshold gain g_{th} must compensate losses:

Eq. (6.42) $$\Gamma g_{th} = \alpha_i + \alpha_m$$

where Γ is the confinement factor which is the fraction of stimulated output mode power guided by the active region.

For a current density J, the carrier density rate equation is:

Eq. (6.43) $$\frac{\partial n}{\partial t} = D(\nabla^2 n) + \frac{J}{qd} - \frac{n}{\tau}$$

where n denotes the carrier density, d is the thickness of the active region, and τ is the lifetime of the non-equilibrium carriers. The first term of the above equation accounts for carrier diffusion with a diffusion coefficient D. The second term governs the rate at which the carriers are injected into the active layer. Since the active region dimensions are usually much smaller than the diffusion length, we assume the carrier density does not vary significantly over the active region so that the diffusion term can be neglected. Therefore, from $\frac{\partial n}{\partial t} = 0$ at steady-state, we get:

Eq. (6.44) $$J = \frac{qnd}{\tau}$$

When the threshold condition is reached, the carrier density is pinned at threshold value n_{th}, and the threshold current density can be expressed as:

Eq. (6.45) $$J_{th} = \frac{qn_{th}d}{\tau}$$

An important factor which determines laser output power is the internal quantum efficiency η_i which is the percentage of the injected carriers that contribute to radiative transitions. So, in the cavity, the photon density can be written as:

Eq. (6.46) $$N_{pn} = \eta_i \frac{J - J_{th}}{qd} \tau_p$$

where τ_p is the photon lifetime which is defined by $\tau_p^{-1} = v_g(\alpha_m + \alpha_i)$. $v_g = c/\overline{n}$ is the group velocity of the light. Since photons escape out of the cavity at a rate of $v_g\alpha_m$, the output power is related to the photon density by the relation:

Eq. (6.47) $$P = \hbar\omega \cdot v_g \alpha_m \cdot VN_{ph} = \hbar\omega \cdot v_g \alpha_m \cdot V \cdot \eta_i \frac{J - J_{th}}{qd} \frac{1}{v_g(\alpha_m + \alpha_i)}$$

where V is the volume of the active region. If we neglect current leakage, i.e. we assume that all the injected current passes through the active region, then the current can be written as $I = JS$, where S is the area of the active region. Considering $V = d \cdot S$, the output power depending on driven current $I > I_{th}$ is rewritten as:

Eq. (6.48) $$P = \frac{\hbar\omega}{q}\frac{\alpha_m}{\alpha_m + \alpha_i}\eta_i (I - I_{th})$$

Typical electrical and laser output power characteristics are shown in Fig. 6.18, in which V_t is the turn-on voltage for the diode, R_s is the series resistance above diode turn-on, and η_s is the slope efficiency.

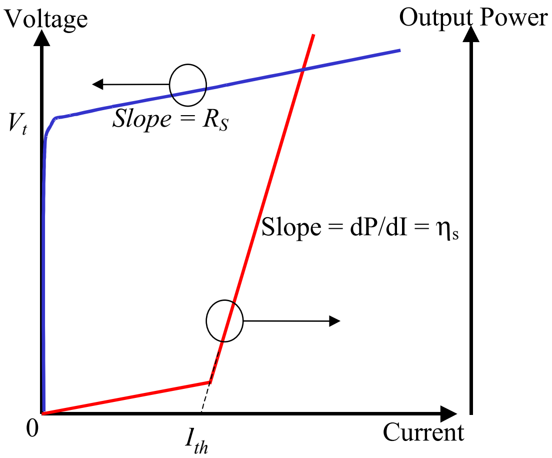


Fig. 6.18. Typical electrical and laser output power characteristic of a semiconductor laser, visualized as the current-voltage and output power-current characteristic. The linear part of the current-voltage curve gives the diode series resistance (R_s), while the linear part of the output power-current curve yields the slope efficiency (η_s).

In real lasers, the linear current dependence in Eq. (6.48) saturates due to such factors as leakage of carriers from the active region, heating, and current induced increases in the internal loss α_i.

The external differential quantum efficiency is defined as:

Eq. (6.49) $$\eta_d = \frac{dP/dI}{\hbar\omega/q} = \eta_i \frac{\alpha_m}{\alpha_i + \alpha_m}$$

Since the optical field should reproduce itself after each round trip under steady-state, the mirror loss α_m can be determined by:

Eq. (6.50) $$R_1 R_2 \exp(-\alpha_m 2L) = 1$$

where R_1, R_2 are the facet reflectivities at the two ends and L is the cavity length. From the above equation we obtain:

Eq. (6.51) $$\alpha_m = \frac{1}{2L}\ln\left(\frac{1}{R_1 R_2}\right)$$

When $L \to 0$, we have $\alpha_m \to \infty$ and $\eta_d \to \eta_i$. Therefore, by plotting $1/\eta_d$ versus L and extrapolating to $L = 0$, the internal quantum efficiency can be determined. Further the slope of the curve is proportional to the internal loss, α_i. In III-V double-heterostructure lasers η_i is close to unity and α_i ranges from 1~100 cm^{-1}.

6.5.3. Linewidth of semiconductor laser diodes

The linewidth in laser diodes depends on the instantaneous changes of phase and intensity in the lasing field [Henry et al. 1982]. These instantaneous changes of phase and intensity have two components: (1) The components directly related to the spontaneous emission and (2) The components directly related to the coupled relationship between the phase and intensity of the field. These changes in field induce a perturbation in the carrier distribution (affecting the real and imaginary parts of the refractive index) which is incorporated into equation by a linewidth enhancement factor denoted as α. The mathematical derivation is rather lengthy and the reader is referred to the book by Chuang [1995]. The linewidth Δf of a typical laser in terms of frequency is then given as:

Eq. (6.52) $$\Delta f = \frac{v_g^2 h\upsilon g n_{sp} \alpha_m (1+\alpha^2)}{8\pi P}$$

where n_{sp} is the spontaneous emission factor which is related to the spontaneous emission rate R_{sp} via $R_{sp} = v_g g n_{sp}$.

6.5.4. Homojunction lasers

The first semiconductor laser was realized using a simple *p-n* junction as shown in Fig. 6.16. This is referred to as a homojunction, i.e. using the same semiconductor material for the active and surrounding layers. In this case, the difference in refractive index between the active layer and the adjacent layers is only 0.1~1%. The result of this is that a lot of the emitted light escapes without undergoing feedback and amplification.

The primary benefits of the homojunction laser rely on simplicity of design and compactness compared to gas and traditional solid state lasers. The low confinement factor and high absorption loss lead to a very high threshold current density at room temperature ($>100\ \text{kA}\cdot\text{cm}^{-2}$) and low power conversion efficiency. These problems are solved in part by making use of a more elegant design, the heterojunction laser.

6.5.5. Heterojunction lasers

To obtain more efficient lasers, it is necessary to use multiple layers with different optical properties in the laser structure. When dissimilar materials are combined, a heterojunction laser can be formed.

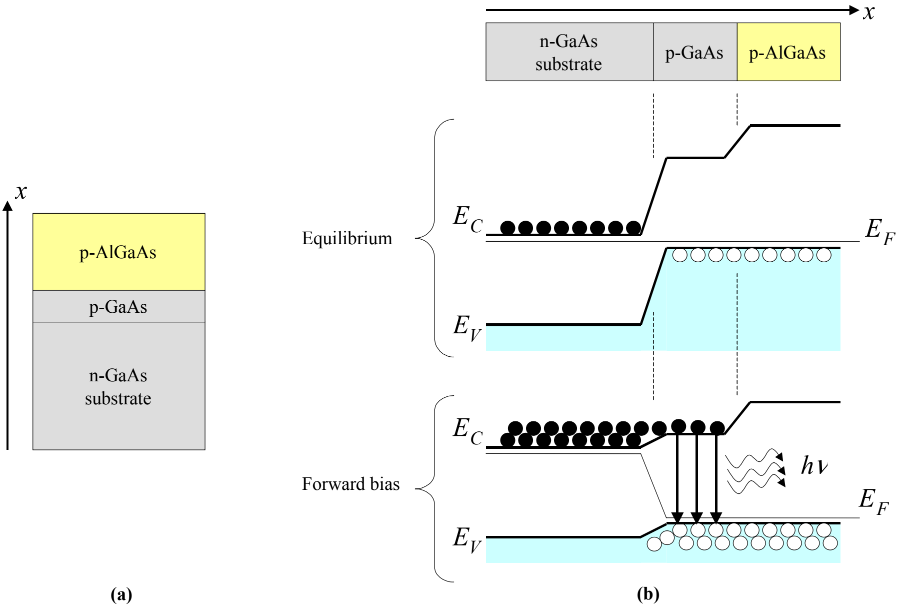


Fig. 6.19. Illustration of the use of a single heterojunction for carrier confinement in laser diodes: (a) cross-section schematic of an AlGaAs heterojunction grown on a thin p-type GaAs layer and on an n-type GaAs substrate; (b) energy band diagrams for this structure at equilibrium and under high forward bias, showing the confinement of electrons into the thin p-type region under forward bias.

An example of a single heterojunction laser is shown in Fig. 6.19. Carrier confinement is obtained in this single-heterojunction laser by using an AlGaAs layer grown epitaxially on GaAs. In this structure the injected carriers are confined to a narrow region so that population inversion can be built up at lower current levels. Further, because there is a noticeable refractive index change at the GaAs/AlGaAs interface, some waveguiding inside the epilayer and substrate is possible. These two effects help to reduce the average threshold current density at room temperature to ~10 kA·cm^{-2}.

A further improvement can be obtained by sandwiching the active GaAs layer between two AlGaAs layers. This double-heterojunction (DH) structure further confines injected carriers to the active region, and refractive index steps at the GaAs-AlGaAs boundaries form the waveguide that confines the generated light waves. A double-heterojunction laser, also called double-heterostructure laser, is shown in Fig. 6.20.

To date, the most extensively used heterostructure lasers are in the GaAs-AlGaAs and GaAs-InGaAsP systems. The ternary alloy Al$_x$Ga$_{1-x}$As has a direct bandgap for x up to $x \approx 0.45$, then becomes an indirect bandgap semiconductor. For heterostructure lasers, the composition region $0 < x < 0.35$ is of most interest and the direct energy gap of the ternary compound can be expressed as:

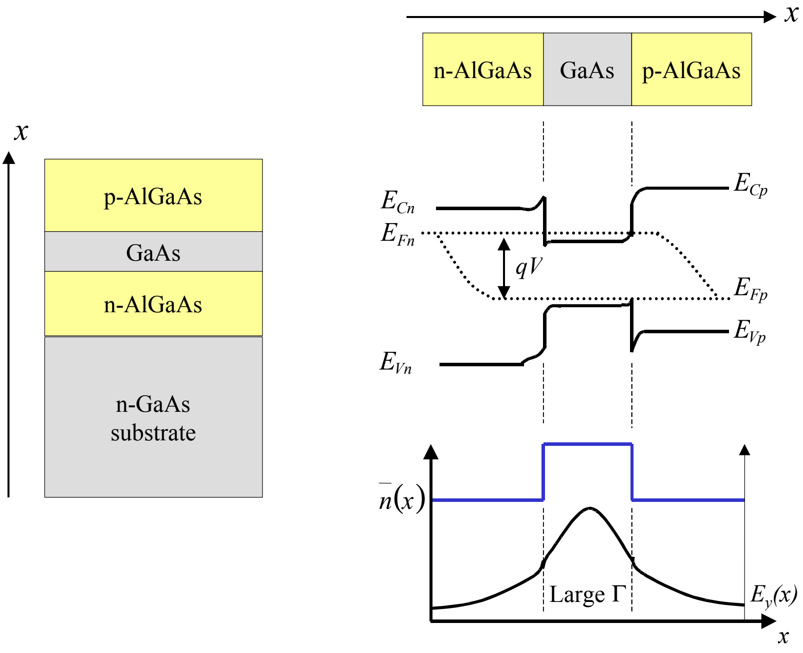
Eq. (6.53) $E_g = 1.424 + 1.247x$ (eV)

The compositional dependence of the refractive index can be represented by:

Eq. (6.54) $\bar{n}(x) = 3.59 - 0.71x + 0.091x^2$

For example, for $x = 0.3$ the bandgap of Al$_{0.3}$Ga$_{0.7}$As is 1.798 eV which is 0.374 eV larger than GaAs; its refractive index 3.385 is about 6% smaller than the GaAs.

Because a DH laser requires several different materials and a controlled doping profile, more sophisticated epitaxial techniques had to be developed, as described in Chapter 1. The deposition can be based on liquid, vapor, or atomic beam processes. All processes allow some degree of multilayer growth with accurate *n*- and *p*-type doping done in-situ. Lastly, in order to avoid threading dislocations, the various crystal layers should be lattice-matched to the substrate.



(a) **(b)**

Fig. 6.20. Illustration of a double-heterojunction laser structure used to confine injected carriers and provide waveguiding for the light: (a) an n-type doped AlGaAs layer has been added between the n-type GaAs substrate and the GaAs active layer in the structure of Fig. 6.19 to form the double-heterostructure; (b) band structure and optical waveguide properties of the resulting laser structure at forward bias (V). The Solid lines in the band structure represent band edges, while the dashed lines represent quasi-Fermi levels.

The optical and electrical confinement of the double heterostructure give it a significant advantage over the homojunction and single-heterojunction laser. Indeed, the threshold current density is reduced by an order of magnitude for most structures. However, in order to maintain a high confinement factor and minimize loss in the cladding regions, the active region thickness must be quite large (0.1 ~ 0.5 μm). The thick layer, combined with a large density of states in the active region, requires a large number of carriers to maintain a sufficient population inversion. Modern DH lasers have thresholds current densities ~1 kA·cm^{-2}.

6.5.6. Device fabrication

After the laser structure is designed and epitaxially grown, a laser device must be fabricated using the photolithographic and metallization processes discussed in Chapter 3. Metal contacts are used to inject electrons and holes into the active region, therefore allowing the creation of a population inversion. In order to realize a low contact resistance between the metal and the semiconductor film, most laser structures employ a highly doped cap layer directly on top of the waveguide cladding layers. This will tend to form an ohmic rather than Schottky contact when metal is evaporated onto the surface.

The other feature a laser must exhibit is optical feedback. After the contact region is formed, the laser must be diced into separate cavities for testing. As discussed earlier, this is often accomplished by taking advantage of the natural cleavage planes in the crystal material. Cleaving is most easily achieved when the substrate is thinned down.

Subsequently, for testing and excess heat removal, the laser chip must be mounted on some type of heat sink. The heat sink also supplies mechanical stability, while allowing electrical connection to an outside circuit.

Example 1. Broad area laser fabrication

The simplest semiconductor laser to fabricate is a broad area laser. The typical fabrication steps for such a laser are shown in Fig. 6.21. No lithography is needed and the fabrication is just enough to satisfy the above requirements. A thin layer of metal is evaporated or electrochemically deposited on the top surface. The type of metal used depends on the material and doping type of the semiconductor, and is chosen to yield a low resistance, ohmic metal-semiconductor contact. The metal is typically annealed to both increase the metal adhesion and establish the ohmic contact.

In order to decrease electrical and thermal resistance, as well as make the wafer more conducive to cleaving of Fabry-Perot cavities, the back side of the wafer is thinned. Lapping and polishing brings the wafer thickness down to ~100 μm. The polished surface is then cleaned and the bottom metal contact is deposited and annealed, as shown in Fig. 6.21(c) and (d). After both contacts are formed, the wafer is cleaved into individual laser cavities with a well-defined length and width, as shown in Fig. 6.21(e).

After cavity formation, the laser is then die bonded to a submount, or directly onto a heat sink (Fig. 6.21(f)). Besides removing waste heat, the heat sink provides mechanical stability and allows macroscopic external electrical contact.

Broad area lasers use one of the fastest fabrication methods. The device features confinement of light and carriers within the large slab formed by the cleaving process. Unfortunately, mechanical cleaving can only reliably produce lengths and widths >200 μm. Further, the cleaving produces minor damage where the surface was scribed along the lateral edges. The net effect is a large current requirement and an increased chance of failure.

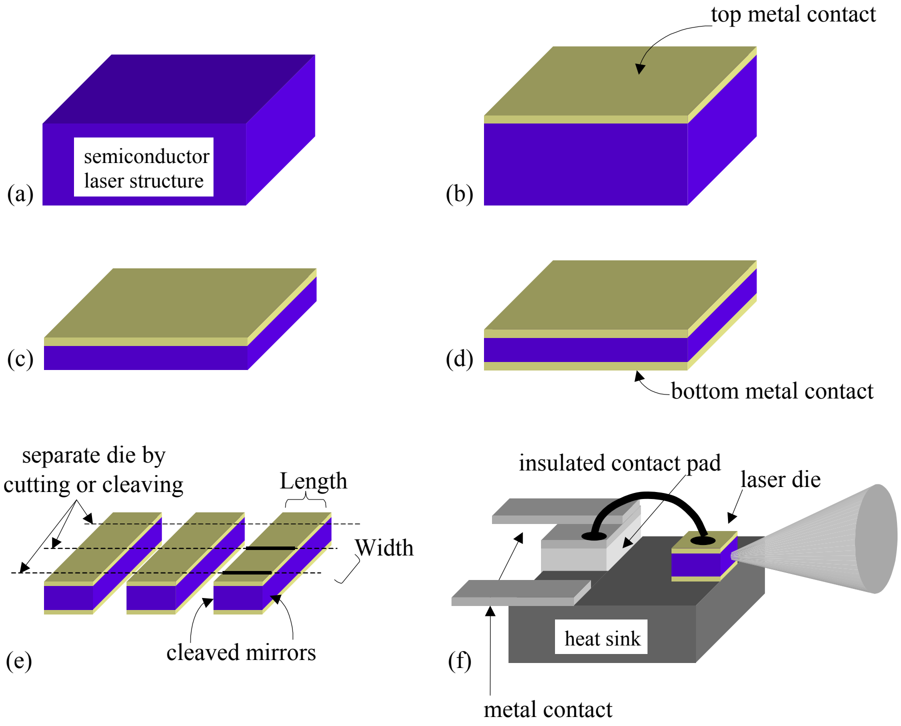

top metal contact
semiconductor laser structure
(a)
(b)
(c)
(d)
bottom metal contact
separate die by cutting or cleaving
Length
Width
cleaved mirrors
(e)
insulated contact pad
laser die
heat sink
(f)
metal contact

Fig. 6.21. Broad area laser fabrication steps: (a) bare semiconductor laser structure, (b) top contact metallization, (c) substrate lapping and polishing to thin down the substrate, (d) bottom contact metallization, (e) cleaving of laser cavities, (f) bonding of laser die to heat sink and external contact formation.

Example 2. Stripe-geometry laser fabrication

Stripe-geometry lasers are lasers in which the current is restricted along the junction plane. In this technique, metal contact stripes are defined which are typically 5~200 μm wide. This technique, while more difficult to fabricate, does allow for a lower operating current and reduced failure rate by keeping the injected area small and far from the (lateral) edges of the chip.

Stripe-geometry lasers can be fabricated in a variety of ways. One of the simplest techniques is shown in Fig. 6.22. The cavity width is defined by a patterned insulator, such as SiO_2. The insulator is typically deposited using chemical vapor deposition, RF sputtering, or e-beam evaporation. Patterning is done using standard optical photolithography and etching, as described in Chapter 3.

The etching of the insulator can typically be done with chemicals or plasma. For SiO_2, a buffered hydrofluoric acid (HF) solution is generally a good selective chemical etchant. The plasma process varies, but typically uses CF_4 or a similar fluorocarbon species.

After the stripe definition, the process followed is similar to the broad area laser. The top contact metallization is followed by substrate thinning and bottom contact metallization. As shown in Fig. 6.22(d), the current injection can be well confined to the area under the insulator opening. Laser die bonding is straightforward, and device operation is relatively insensitive to the lateral edges of the chip.

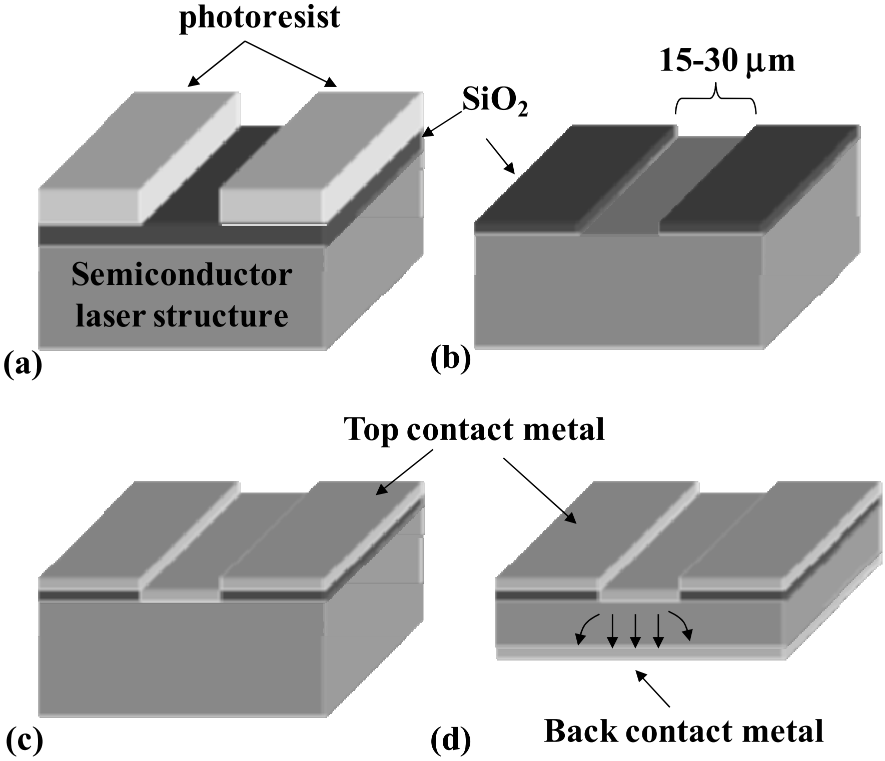


Fig. 6.22. Stripe-geometry laser fabrication steps and schematic: (a) photolithography on SiO_2 insulation, (b) patterned SiO_2, (c) top contact metallization, (d) schematic of final device showing localized current injection paths.

One of the disadvantages of the stripe-geometry laser is lateral current leakage. Even though the width of the cavity is defined at the surface, the injected current spreads out as it travels toward the substrate. The relative amount of spreading at a given current depends on the stripe width as well as the lateral conductivity and carrier diffusion length of the layers.

Because the gain of the laser varies along the junction plane due to this current spreading, the effective complex refractive index is also non-uniform. The complex refractive index is highest at the center of the stripe and decreases in a quadratic manner with distance, until it reaches its equilibrium value. A weak waveguiding effect is anticipated, as derived in Casey and Panish [1978]. This is referred to as a gain-guided laser.

While helpful in reducing the threshold current requirements, gain-guiding typically shows multiple transverse mode output in the junction plane.

Example 3. Buried-heterostructure laser

The last structure that will be discussed is referred to as a buried-heterostructure laser. This device goes one step further in complexity in order to completely confine both the current and the optical mode around a small emitting core.

The fabrication of the buried-heterostructure lasers, as shown in Fig. 6.23, starts with the growth of the *n*-type waveguide cladding and active layer(s). Using photolithography, the core is patterned with photolithography and etching into very narrow stripes (Fig. 6.23(b)). The width is small in order to confine only a single transverse optical mode, and depends on the laser emission wavelength as well as the index difference between the core and cladding regions.

After the core is defined, epitaxial regrowth is used to cover the core with a low index, high-bandgap, *p*-type cladding (Fig. 6.23(c)). Following this step, standard thinning, metallization, and die bonding is performed to complete the device.

While epitaxial regrowth is technologically very challenging, this fabrication procedure has the potential of producing the most efficient lasers with a single transverse mode. The current is confined thanks to the cladding-core material band offset, and the light is confined thanks to the refractive index change. The output beam quality is generally very high, which makes these lasers attractive for fiber coupling and telecommunications. The only drawback is a low output power, which scales as the core volume.

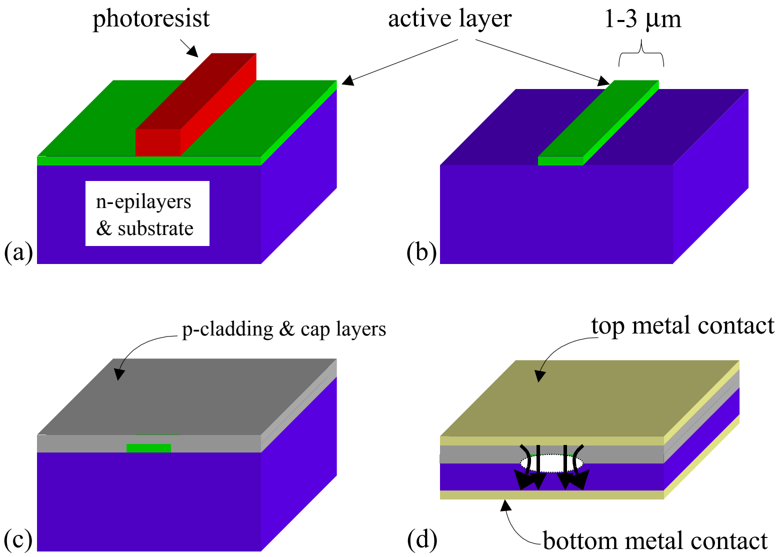
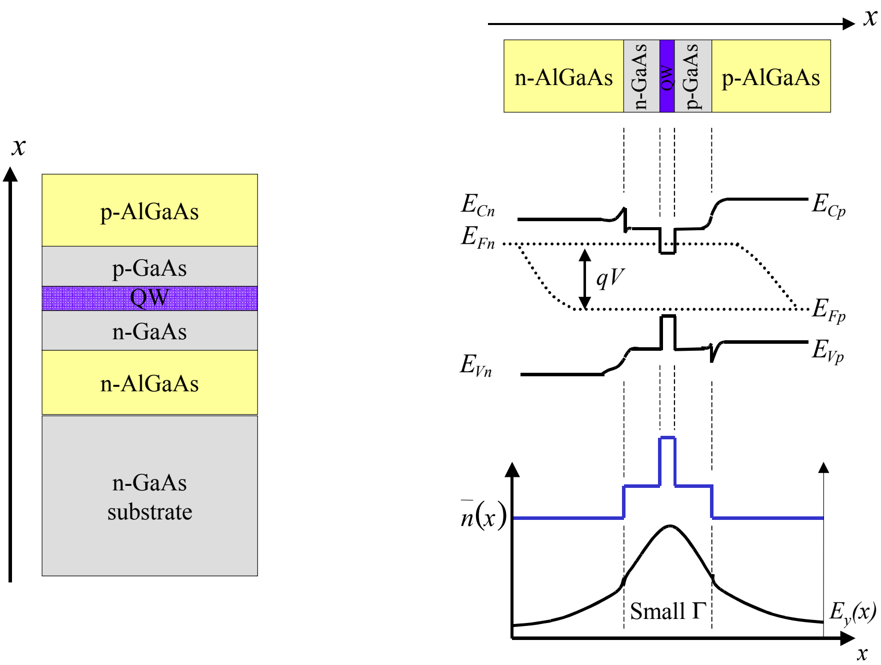

photoresist
active layer
1-3 μm
n-epilayers
& substrate
(a)
(b)
p-cladding & cap layers
top metal contact
(c)
(d)
bottom metal contact

Fig. 6.23. Fabrication steps and schematic of buried-heterostructure laser: (a) photolithography to define waveguide core, (b) patterning of core, (c) semiconductor regrowth of waveguide cladding, (d) schematic of final device showing the confined current and optical confinement to achieve single transverse mode output.

6.5.7. Separate confinement and quantum well lasers

A further advancement on the double heterostructure laser uses several different kinds of materials to separate the optical and electrical confinement into separate regions. The separate confinement heterostructure (SCH) typically uses a thin or quantum well-based active region surrounded by an intermediate waveguide layer, all of which are embedded in the standard high-bandgap cladding region. A schematic of the heterostructure and the optical waveguide properties under forward bias are shown in Fig. 6.24.


x
p-AlGaAs
p-GaAs
QW
n-GaAs
n-AlGaAs
n-GaAs substrate
x
n-AlGaAs
n-GaAs
QW
p-GaAs
p-AlGaAs
E_{Cn}
E_{Fn}
qV
E_{Cp}
E_{Fp}
E_{Vn}
E_{Vp}
$\bar{n}(x)$
Small Γ
$E_y(x)$
x

(a) (b)

Fig. 6.24. Illustration of a separate confinement heterostructure laser: (a) cross-section of the device structure; (b) band structure, index and optical mode profiles of the resulting laser structure at forward bias (V). The Solid lines in the band structure represent band edges, while the dashed lines represent quasi-Fermi levels.

For quantum well (QW) or multi-quantum well (MQW) active regions, the density of states is reduced. This fact, combined with a narrow width for the well (<30 nm) leads to a population inversion at very low current densities. The gain also takes on another general form as a function of carrier density and is given by:

Eq. (6.55) $$g_w = g_0\left[\ln\left(\frac{J_w}{J_0}\right)+1\right]$$

where g_0 and J_0 are the transparency gain and current density respectively. Furthermore, because the carrier injection is more uniform, the internal quantum efficiency can be higher. The inserted waveguide layer helps distribute the optical mode in order to reduce the divergence of the laser beam. In addition, as the waveguide layer is nominally undoped, free-carrier absorption is reduced.

Unfortunately, the confinement factor is rather small for the active region (1~5%). Despite this apparent disadvantage, for a low modal threshold gain, defined as:

Eq. (6.56) $G_{th}=\Gamma g_{th}$,

the separate confinement heterostructure (SCH) can still be designed to reach threshold significantly earlier than a double heterostructure, as shown schematically in Fig. 6.25.

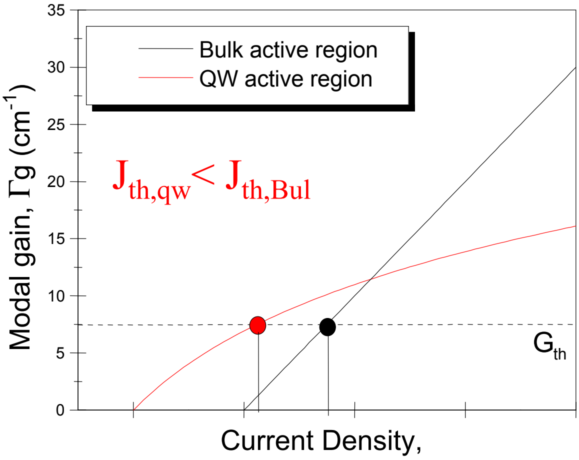


Fig. 6.25. Schematic comparison of QW and DH gain behavior as a function of current density. For a low threshold modal gain, the SCH can have a significantly lower threshold current density, thanks partially to the reduced density of states in the active region.

Another benefit of using thin and/or QW active regions is the possibility to realize strained layers. Indeed, unlike the double heterostructure (DH) laser where all materials needed to be lattice-matched, the SCH can incorporate strained materials (such as $In_xGa_{1-x}As$ on GaAs) as long as the layer is thinner than the critical thickness above which threading dislocations start to form. Strained layers can have two important benefits. The first improvement has to do with the band structure of a strained semiconductor. For compressively strained material, the heavy-hole mass becomes lighter than bulk, which leads to a reduced hole density of states, and further reduction of the transparency current density.

Finally, the accessible wavelength range for a given substrate is increased. For example, a DH laser based on GaAs has a maximum emission wavelength of 870 nm. Using strained quantum wells, a SCH laser can extend the emission wavelength well past 1 µm.

6.5.8. Laser packaging

Regardless of the fabrication procedure, the packaging of the laser diode depends on the final application.

The simplest package consists of an open heat sink, similar to the type used in the broad area laser fabrication (Fig. 6.21(f)). These come in many different designs, depending on the size restrictions in the final product. One example of an open package is shown in Fig. 6.26(a). Another popular product, which also protects the laser, is a transistor-type can with an output window, as shown in Fig. 6.26(b). These typically are sold in 5 and 9 mm diameter sizes as well as a larger TO-3 package. For high power lasers or laser bar packaging, heat management is very important. In this case, a larger mass submodule is used, which frequently incorporates a thermoelectric cooler (TEC) and controller to keep the temperature stable. A representative high heat load package, is shown in Fig. 6.26(c). Other applications benefit from having the laser coupled directly to a fiber optic cable. In this case, the fiber is carefully aligned and welded into place inside a hermetically sealed package, as shown in Fig. 6.26(d).

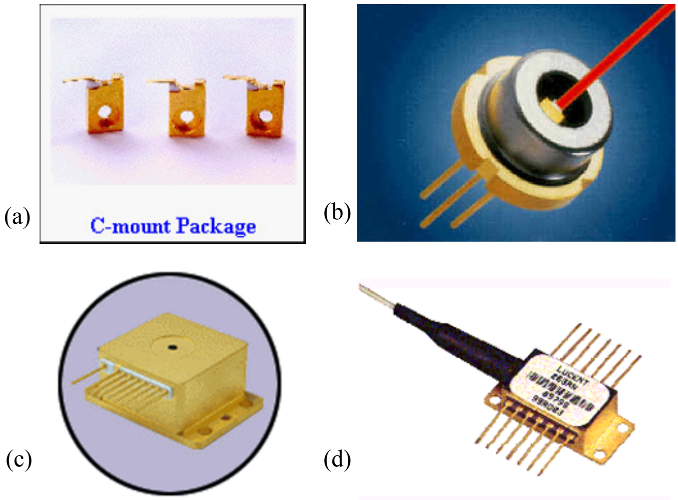


Fig. 6.26. Examples of commercial laser package designs. (a) open "c-mount" type heat sinks; (b) 9 mm transistor can package; (c) high heat load package incorporating laser and cooler; (d) fiber-coupled "butterfly" package.

6.5.9. Distributed feedback lasers

Semiconductor lasers typically exhibit multiple wavelength emission at high current because of the presence of a Fabry-Perot cavity, as illustrated in Fig. 6.27(b). This occurs when multiple longitudinal modes reach the laser threshold gain. When only one wavelength is desired, a common technique is to realize a secondary feedback within the cavity. A corrugated grating positioned inside the waveguide is one means to this end, as shown in Fig. 6.27(a). This type of laser is referred to as a distributed feedback (DFB) laser, because the feedback effect is distributed over most, if not all, of the laser cavity.

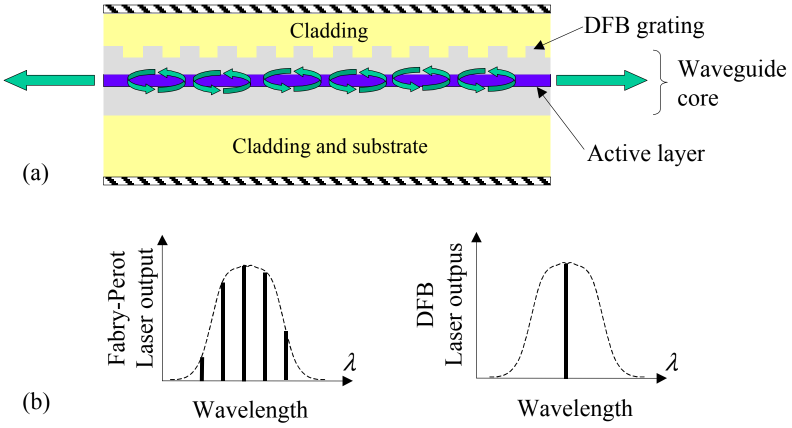


Fig. 6.27. (a) Schematic of the cross-section of a DFB laser. Periodic variations in the effective refractive index leads to distributed optical feedback; (b) proper grating design can change the output spectrum of a laser structure from multi- to single-wavelength.

The grating geometrical parameters are chosen to satisfy Bragg's law of diffraction:

$$m\frac{\lambda_0}{\overline{n}_{eff}} = 2\Lambda\sin(\theta) \qquad \text{Eq. (6.57)}$$

where m is the grating order, λ_0 is the free space wavelength, $\overline{n}_{eff}$ is the effective refractive index in the waveguide, Λ is the grating period, and θ is the diffraction angle. The minimum requirement for optical feedback is diffraction at 180° relative to the propagation of the waveguide mode, i.e. the diffraction angle is $\theta = 90°$. For a first order grating ($m = 1$), Bragg's law is fulfilled only when:

$$\Lambda = \frac{\lambda_0}{2\overline{n}_{eff}} \qquad \text{Eq. (6.58)}$$

The diffraction of light by the grating serves to enhance the cavity reflectivity at the designed wavelength. The threshold gain is then reduced for this wavelength, allowing favored laser oscillation of a specific longitudinal mode, in some cases (Fig. 6.27(b)).

Unfortunately, for many near-infrared lasers, the period of a first order grating is in the 100–300 nm range, a resolution which requires a high-end lithography setup. To make fabrication easier, second, third, and fourth order gratings have also been explored for this application, though their efficiency is somewhat lower due to lower order diffraction loss.

Despite this technological difficulty, DFB lasers are frequently incorporated into buried-heterostructure designs in order to demonstrate true single mode (longitudinal and transverse) behavior. This is especially useful in telecommunications and chemical spectroscopy, where a stable, monochromatic laser (single longitudinal mode wavelength) is preferred.

6.5.10. Material choices for common interband lasers

Table 6.1 summarizes some of the most common III-V material systems used in various parts of a laser structure in order to achieve a specific range of laser emission. The parts of the laser considered include the substrate, cladding and active region materials.

Substrate	Cladding material	Active Material	Wavelength Range
GaN(*)	$Al_xGa_{1-x}N$	strained $In_xGa_{1-x}N$ QWs	400~600 nm
GaAs	$Al_xGa_yIn_{1-x-y}P$	$Ga_{0.51}In_{0.49}P$	660 nm
GaAs	$Al_xGa_{1-x}As$ or $Ga_{0.51}In_{0.49}P$	$Ga_xIn_{1-x}As_yP_{1-y}$	660~870 nm
GaAs	$Al_xGa_{1-x}As$ or $Ga_{0.51}In_{0.49}P$	strained $In_xGa_{1-x}As$ QWs	0.87~1.1 μm
InP	$Al_{0.48}In_{0.52}As$	InP	920 nm
InP	InP	$Ga_xIn_{1-x}As_yP_{1-y}$	1~1.7 μm
InP	InP	$Ga_{0.47}In_{0.53}As$ QWs	1.3~1.7 μm
InP	InP	strained $In_xGa_{1-x}As$ QWs	1.5~2 μm
InAs	$Al_xGa_yIn_{1-x-y}As_zSb_{1-z}$ $InAs_xSb_yP_{1-x-y}$	InAs QW	2~3 μm
InAs	$Al_xGa_yIn_{1-x-y}As_zSb_{1-z}$ $InAs_xSb_yP_{1-x-y}$	strained $InAs_ySb_{1-y}$ QWs	3~5 μm

Table 6.1. Materials for substrates, cladding and active regions in semiconductor lasers for various emission wavelength ranges. (*) Not a mature substrate technology. Lasers are often grown on sapphire or SiC substrates.

6.5.11. Interband lasers

GaAs-based lasers in general have found many applications from simple laser pointers to CD/DVD players. In addition, significant development has also been invested in achieving very high power efficiency and output powers from these devices for applications such as welding/cutting, frequency doubling, and solid-state laser pumping.

The last application especially has drawn a lot of interest, as diode-pumped laser systems show significantly higher efficiency and lifetime compared to conventional flashlamp-pumped systems. The primary reason for this is that the diode laser emission wavelength can be closely matched to the narrow absorption peak of the solid-state laser medium. This allows most of the emitted radiation from the laser diode to be directly absorbed. Flashlamps, on the other hand, are broadband sources, and a large fraction of the light goes directly to waste heat. Further, flashlamps have a limited lifetime, on the order of five hundred to several thousand hours. On the other hand, the durability of aluminum-free diode lasers was tested by operating them under continuous wave at 1 W, 60 °C for an extended period of time. They exhibited no degradation over 30,000 hours [Diaz et al. 1996] [Diaz et al. 1997]. Under normal operating conditions (20 °C) the projected lifetimes are on the order of several million hours.

The aluminum-free GaInAsP technology was used to achieve high performance semiconductor lasers emitting at 980 nm [Mobarhan et al. 1992] through the use of strained InGaAs quantum wells inside a separate confinement heterostructure (SCH). These lasers exhibited a low J_{th}~70 A/cm^2, high differential efficiency ~1.0 W/A, and low internal loss of 1.5 cm^{-1}. They also yielded high output power and a very high characteristic temperature T_0, over 350 K in the range of 20~40 °C. These 980 nm lasers could operate under continuous wave at high power (1.4 W) at 100 °C.

Optimized 808 nm laser diodes based on GaAs/GaInAsP with uncoated facets emitted high output powers of 10 W and 7 W in pulse and continuous wave operation, respectively. Laser bars yielded output powers of 70 W in quasi-continuous wave operation [Razeghi 1994] [Yi et al. 1995]. A properly designed GaInAsP laser structure provides a narrow transverse beam with a divergence of only 26°, which is convenient for efficient laser light coupling into the optical fiber or the pumped Nd:YAG crystal. For comparison, 32–48° are typical values of beam divergence for commercial AlGaAs lasers.

InP based laser have also received a lot of attention thanks to the accessible wavelength range of lattice-matched quaternaries (0.92~1.65 μm). This range makes the system ideal for fiber optic communications, which relies on lasers designed for low-loss and low-dispersion fiber propagation.

Before these lasers could be considered for use in any application, however, several demonstrations had to be made. These initial demonstrations include:

- Double heterostructure $\lambda = 1.3$ µm and 1.55 µm lasers with threshold current densities of 430 and 500 A/cm^2 respectively [Razeghi et al. 1983a] [Razeghi et al. 1983b].
- Buried ridge 1.3 and 1.55 µm lasers with threshold currents as low as 6 mA [Razeghi 1985a].
- Buried ridge lasers with distributed feedback for single wavelength emission (Fig. 6.28) [Razeghi et al. 1985b] [Razeghi et al. 1985c].
- High power phase-locked arrays of 1.3 µm lasers with 600 mW output [Razeghi 1987].
- Separate confinement heterostructure GaInAsP/GaInAs/InP waveguide for improved performance [Razeghi et al. 1985c].

Many other groups took advantage of this technology to produce advanced lasers for telecommunications. [Kuznetsov et al. 1989] With the groundwork laid by these achievements, GaInAsP/InP lasers are now in mass production for the telecommunication industry. There is still active research to extend the wavelength to 2 µm and beyond using strained layer epitaxy.

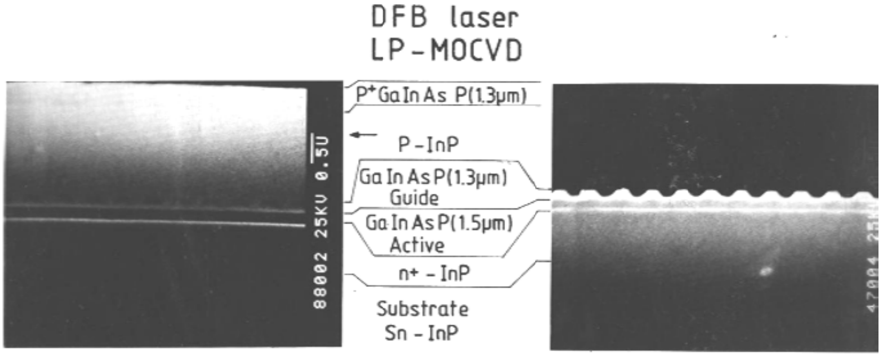


Fig. 6.28. Scanning electron microscope cross-sections of the corrugated structures after and before LP-MOCVD regrowth.

In general, the quest for longer wavelength ($\lambda > 2$ µm) diode lasers has been difficult. There are many applications for longer wavelength lasers, including infrared spectroscopy, infrared countermeasures, free-space communication, and low-loss fiber communication. The only commercially available semiconductor alternative, up until a few years ago, were actually based on IV-VI materials, commonly referred to as lead-salt lasers. This technology has several liabilities, not the least of which are low operating temperature (80 K) and low output power (~1 mW).

Clearly, for most applications it is advantageous to overcome one or both of these liabilities. Unfortunately, increased losses and reduced radiative efficiency are a fact of life for low bandgap zinc-blende semiconductors as well. Despite this challenge, extensive work has been done on InAs-based laser diodes at cryogenic temperatures. This technology has the potential for significantly higher output power powers and is based on an intrinsically more robust material system.

Double heterostructure (DH) lasers based on the InAs/InAsSb/InAsSbP material system were grown by low-pressure Metal-Organic Chemical Vapor Deposition (LP-MOCVD) for high power lasers emitting with $3.1 \leq \lambda \leq 3.4$ µm. A novel asymmetric heterostructure was employed, as shown in Fig. 6.29, which allowed for reduced electron leakage and improved performance. Laser bars consisting of four 100 µm stripes exhibited a maximum peak output power of 6.7 W [Wu et al. 1999]. This registers the highest output power in this wavelength range. Furthermore, optimizing the efficiency of these lasers allowed over 450 mW to be obtained under continuous operation [Razeghi 1998] [Razeghi et al. 1999].

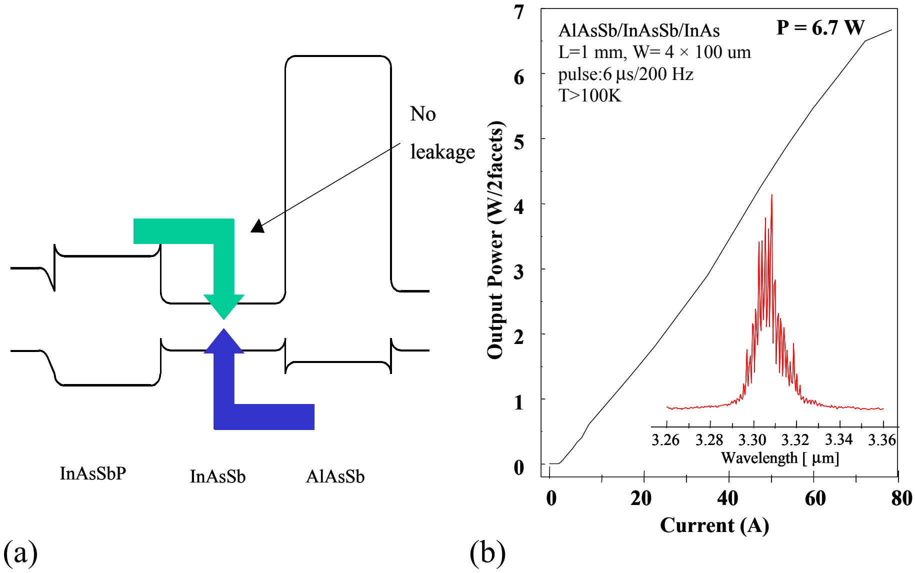


Fig. 6.29. (a) InAsSbP/InAsSb/AlAsSb double heterostructure. (b) P-I curve of InAsSbP/InAsSb/AlAsSb laser bar at T > 80 K.

The first mid-infrared superlattice injection lasers have been designed with InAs/InAsSb, InAsP/InAsSb, and InAsSb/InAsSb strained layer superlattices (SLS). Although complex, SLS lasers benefit from better optical confinement and more emission wavelength flexibility than a DH laser and a larger gain region than a MQW laser. Another advantage that these superlattice lasers have is the reduction of non-radiative recombination (Auger) mechanisms. The first lasers based on the superlattice active region were designed and fabricated for emission from 4~4.8 μm [Lane et al. 1999] [Lane et al. 2000]. This is the longest reported emission from electrically injected lasers based on InAsSb interband transitions.

6.5.12. Quantum cascade lasers

In conventional (interband) semiconductor lasers, light is generated from the recombination of electrons and holes separated by the energy gap. The energy of the photons, or the wavelength of laser light is therefore essentially determined by the energy gap of the semiconductors used. This means, however, that only a certain range of wavelengths can be obtained, which correspond to the energy gap of semiconductors that exist. It turns out that it is quite difficult to have low energy gap semiconductors ($E_g < 0.3$ eV) with high enough quality suitable for semiconductor lasers, and thus it is difficult to make lasers with low photon energy which corresponds to infrared wavelengths ($\lambda > 3$ μm). Because of this limitation, totally different types of semiconductor lasers have been recently fabricated. The first one is the quantum cascade laser, and the other is the type II superlattice laser which will be discussed in the next sub-section. Here, we will briefly describe the operating principle and underlying basic physics of these lasers.

Unlike the interband lasers which have bipolar device characteristics, relying on electron-hole recombination to generate light, the quantum cascade laser (QCL) is a unipolar device which uses only conduction-band electrons. This means that the electron is the dominant carrier and that emitted light is solely due to electron transitions from upper to lower subbands in the conduction band (semiconductor quantum wells have the property of splitting a bulk band structure into a series of subbands). The benefit of this design is that, using the same material, the emission energy can be varied over a wide range (limited by the conduction band offset between the barrier and well) simply by varying the quantum well widths.

The idea of generating light from intersubband transitions within semiconductor quantum wells was first proposed in 1971 by Kazarinov and Suris [1971]. Since then, many groups have tried to produce devices based on similar models. The first electrically pumped intersubband laser at λ=4.2 μm was demonstrated in 1994 by Faist et al. [1994].

In order to have stimulated emission, there has to be a population inversion. This means that there must be a steady injection of high-energy electrons and steady collection of low energy electrons from the same quantum well region. The principle of operation is based on a multi-quantum well structure in an electric field and is illustrated in Fig. 6.30.

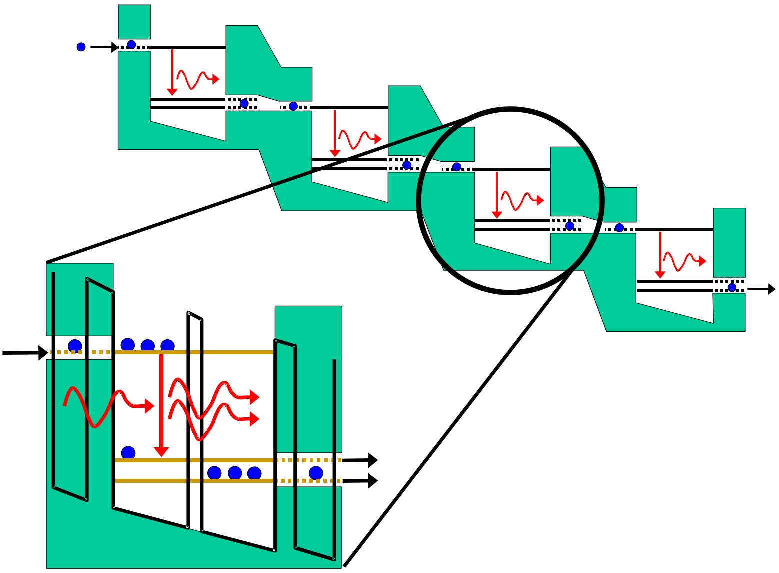


Fig. 6.30. Example of quantum well-based active region for an electrically injected quantum cascade laser. Electrons are injected into the higher energy level in the quantum well, from where they can be stimulated to relax down to the lower energy level and emit a photon. The relaxed electrons can then be transmitted to the next quantum well on the right. This process can be continued by serially stacking active regions for higher power output.

Electrons come to the active region from the injector (left) and go out to a subsequent region (right) which acts as the injector of another active region. All these regions are made from multi-quantum wells. The combination of active and injector regions forms a period of the structure which repeats many times (typically 25~30) forming a cascade structure. This allows a single injected electron to emit multiple photons, which leads to differential efficiencies much greater than unity.

The active region of a typical QCL is a triple AlInAs/GaInAs quantum well structure that supports three subbands as illustrated above. The injector region is made up of the same material and has many wells depending on the emission wavelength. The QCL emission wavelength is controlled solely by the layer thicknesses in the active region, and thus has the potential for a large variety of emission wavelengths in a given material system.

The most important advantage of the quantum cascade laser is its insensitivity to temperature changes. Laser operation is not limited by Auger recombination due to the unipolar nature of the device, which gives a much higher theoretical operating temperature. This high temperature operation will potentially reduce package size and cost due to the absence of elaborate cooling systems. Another significant advantage of QCL is that, as the operation is not directly related to the bandgap of the constituent materials, relatively mature InP or GaAs technology can be used. This ensures that the growth and the physical characteristics of the materials are relatively well understood for QCLs. It also allows for excellent uniformity across 2” wafers, which has already been demonstrated in laboratory situations. Quantum cascade lasers are treated in detail in Chapter 7.

6.5.13. Type II lasers

Type II band alignment and some of its interesting physical behavior were originally suggested by Sai-Halasz et al. [1977]. Soon after that they reported the optical absorption of type II superlattices [Sai-Halasz et al. 1978a] and later the semimetal behavior of the superlattice [Sai-Halasz et al. 1978b]. The applications of such a superlattice was proposed only after several years [Smith et al. 1987]. The flexibility of the material used to cover a huge infrared range (2 to > 50 μm) and the reduced Auger recombination rate [Youngdale et al. 1994] caught the attention of many research groups. Type II heterojunctions have found many applications in electronic devices such as resonant tunneling diodes and hot electron transistors. However, perhaps the most important of these applications has been in the optoelectronics and recently many significant results have been achieved in type II modulator [Xie et al. 1994], detectors [Johnson et al. 1996] [Fuchs et al. 1997], and laser diodes [Yang et al. 1998] [Felix et al. 1997].

Type II lasers are based on active layers which exhibit a type II band alignment. In general, the band alignment of any semiconductor heterojunction can be categorized as type I, type II staggered or type II misaligned, as illustrated in Fig. 6.31. The main difference between type I and type II staggered band alignments resides in the fact that, in the former case, one (same) side of the heterojunction presents a lower energy for both electrons and holes. The electrons and the holes will thus be preferentially found on the same side of the junction. If the bottom of the conduction band of one material is located at a lower energy than the top of the valence band of the other one, we obtain a type II misaligned heterojunction, as shown in Fig. 6.31.

The type of heterojunction depends on the nature of the semiconductors brought in contact. Most semiconductor junctions are of the type I. The special band alignment of the type II heterojunctions provides three important physical phenomena or features which are illustrated in Fig. 6.32(a) a lower effective bandgap, (b) the separation of electrons and holes and (c) tunneling. These properties are used in many devices to improve their overall performance.

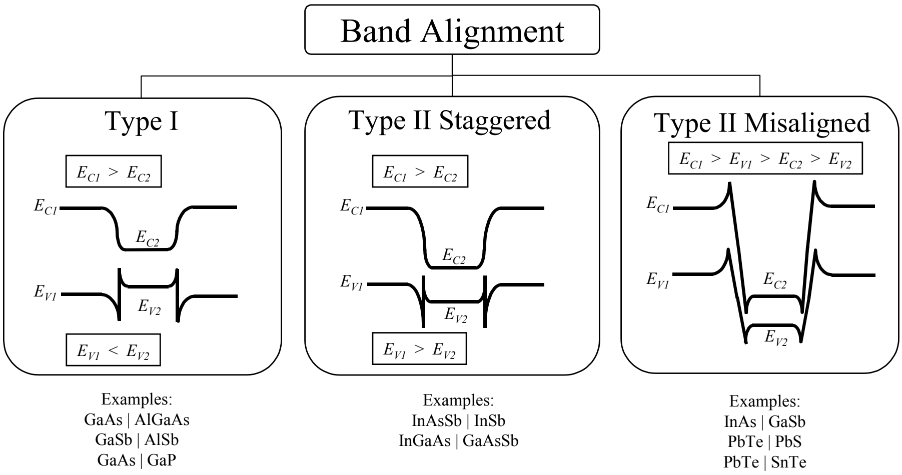

Band Alignment
Type I
$E_{C1} > E_{C2}$
E_{C1}
E_{C2}
E_{V1}
E_{V2}
$E_{V1} < E_{V2}$
Type II Staggered
$E_{C1} > E_{C2}$
E_{C1}
E_{C2}
E_{V1}
E_{V2}
$E_{V1} > E_{V2}$
Type II Misaligned
$E_{C1} > E_{V1} > E_{C2} > E_{V2}$
E_{C1}
E_{V1}
E_{C2}
E_{V2}
Examples:
GaAs | AlGaAs
GaSb | AlSb
GaAs | GaP
Examples:
InAsSb | InSb
InGaAs | GaAsSb
Examples:
InAs | GaSb
PbTe | PbS
PbTe | SnTe

Fig. 6.31. Possible band alignment configurations of a semiconductor heterojunction. The most common band alignment is the type I alignment shown on the left in which one same side of a heterojunction presents a lower energy for both electrons and holes. The type II band alignment is shown in the middle and the right, in which the electrons and holes have a lower energy on different sides of a heterojunction.

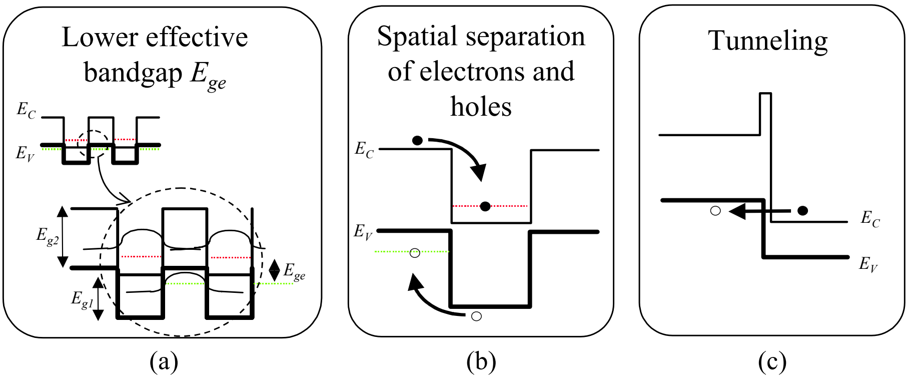

Lower effective bandgap E_{ge}
E_C
E_V
E_{g2}
E_{g1}
E_{ge}
Spatial separation of electrons and holes
E_C
E_V
Tunneling
E_C
E_V
(a)
(b)
(c)

Fig. 6.32. Unique features of type II heterojunctions and superlattices: (a) they have an effective bandgap energy which is lower than those of the constituting semiconductors, (b) electrons and holes are spatially separated, (c) an electron can tunnel from the conduction band on one side of the heterojunction into the valence band on the other side of the heterojunction.

The first feature involves a superlattice with type II band structure. A superlattice is a structure consisting of closely spaced quantum wells, such that the localized discrete energy levels in the quantum wells become delocalized minibands across the entire structure, in both the conduction and the valence bands. These minibands thus form an effective bandgap E_{ge} for the superlattice considered as a whole. Because of the type II band alignment, these minibands can exhibit a lower effective bandgap than the bandgap of either constituting layer, as shown in Fig. 6.32(a), which is a most interesting property. In addition, this effective bandgap is tunable to some extent by changing the thickness of the layers. This is very important for the applications in the mid- and long- wavelength infrared ranges since one can generate an artificial material (the superlattice) with a fixed lattice parameter but with a different bandgap. For example, recently very successful detectors [Mohseni et al. 1997] and lasers [Felix et al. 1997] have been implemented in the 2 to 15 μm wavelength range with InAs/InGaSb superlattices.

The second feature is the spatial separation of the electrons and holes in a type II heterojunction, as shown in Fig. 6.32(b). This phenomenon is a unique feature of this type of band alignment and is due to the separation of the electron and hole potential wells. As a result, a huge internal electrical field exists in the junction without any doping or hydrostatic pressure. High performance optical modulators have been implemented based on this feature [Johnson et al. 1996].

The third feature is the Zener type tunneling of a type II misaligned heterojunction, as shown in Fig. 6.32(c). Electrons can easily tunnel from the conduction band of one layer to the valence band of the other layer, since the energy of the conduction band of the former layer is less than the energy of the valence band of the later layer. However, unlike Zener tunneling, no doping is necessary for such a junction. Therefore, even a semi-metal layer can be implemented with very high electron and hole mobility since the impurity and ion scattering are very low. This feature of type II heterojunctions has been successfully used for resonant tunneling diodes and, recently, for type II unipolar lasers [Lin et al. 1997].

Type II lasers can generally be categorized as bipolar and unipolar lasers. In the bipolar type II lasers, the structure of a conventional III-V heterostructure laser diode is used except that the active layer is a type II superlattice, as shown in Fig. 6.33. The electron-hole recombination occurs between the first conduction miniband and the first heavy-hole miniband. Since the active layer is a type II superlattice, the energy difference between the minibands can be adjusted by changing the thickness of the layers. This is an important advantage since the laser can potentially cover a wide range of wavelengths (from ~2 μm to beyond 50 μm) without changing the chemical composition of the materials in different layers.

In a unipolar type II laser, electrons are injected into the active layer through an injection layer, similar to a quantum cascade laser. They then radiatively recombine with a hole, unlike a quantum cascade laser. The electrons can also tunnel through a type II tunneling junction, and go to the next injection layer and repeat a similar transition in a cascade fashion. The interband transition in these lasers leads to an important advantage compared to the quantum cascade lasers. This type of transition is much more immune to the phonon scattering than the intersubband transition, and hence the efficiency of unipolar type II lasers is much higher than that of quantum cascade lasers.

It should be noted that type II lasers are interband lasers, meaning that it involves the radiative recombination of an electron from the conduction band with a hole in the valence band. In such types of lasers, Auger recombination can generally play an important role and limit the high temperature operation of such infrared lasers. Fortunately in the case of a bipolar type II laser, the spatial separation of electrons and holes and the band structure of the superlattices lead to a much lower Auger recombination rate than in conventional semiconductor infrared lasers. Recently, bipolar type II lasers yielded the highest operating temperature for light emission at 3.2 μm.

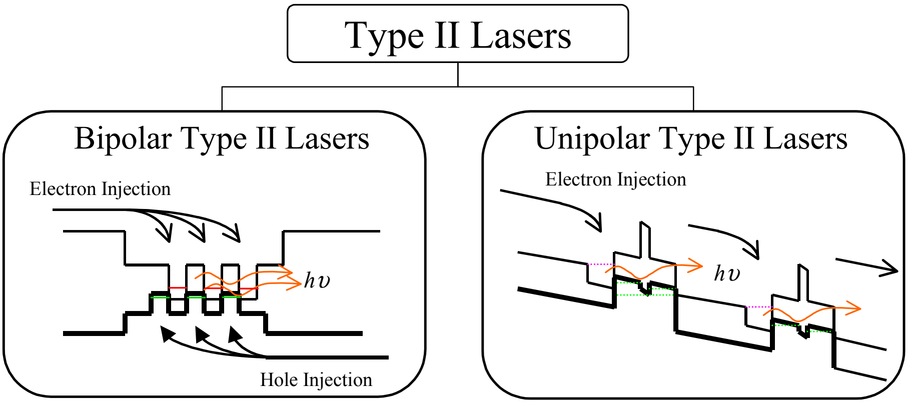


Fig. 6.33. Bipolar and unipolar type II lasers. In a bipolar laser, the electrons and holes are injected and recombine radiatively in the quantum well or superlattice type II structures. In the unipolar type II laser, electrons are injected and emit photons by relaxing into a lower energy state in the heterostructure.

6.5.14. *Vertical cavity surface emitting lasers*

In all the lasers discussed so far the light is emitted from the edges of active region. There has always been a desire for surface-emitting lasers which would allow free-space communication from chip to chip between specific locations and 2D arrays of high-power light sources. Earlier attempts have introduced mirrors or gratings to turn the edge-emitted light by 90°, but recently major advances have been made by using GaAs-AlGaAs or GaAs-AlAs multilayers (grown at the same time as the laser structure) as integral mirrors.

The multilayer mirrors, known as Bragg reflectors, reflect light at certain wavelengths, in particular that of the laser. The GaAs and AlGaAs (or AlAs) layers forming the mirrors are each typically a quarter-wavelength thick (Fig. 6.34).

The structure is shown in Fig. 6.34, together with a photograph of an array of such devices. It is straightforward to place ohmic contacts on the top and the bottom, and there is no need for cleaving and polishing end facets. The three main features that determine their operation are (i) the high reflectivity of the mirrors, (ii) the means of lateral confinement of the active region, and (iii) the limitations imposed by laser heating from carrier injection through the resistive mirrors.

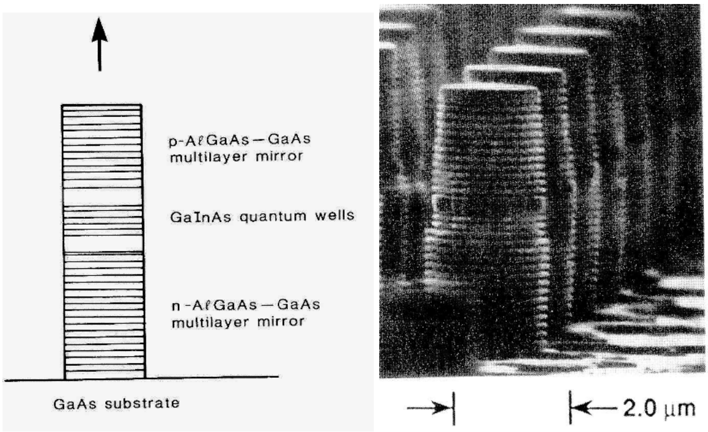


Fig. 6.34. Example of a multilayer structure for a surface-emitting laser: GaInAs quantum wells are sandwiched between two GaAs-AlAs multilayer mirrors, one n-type doped and the other p-type doped. ["Figure 18.10", from Low-Dimensional Semiconductors: Materials, Physics, Technology, Devices by M.J. Kelly; taken after Applied Physics Letters Vol. 55, Scherer, A., Jewell, J.L., Lee, Y.H., Harbison, J.P., and Florez, L.T., "Fabrication of microlasers and microresonator optical switches," p. 2724-2726. Copyright 1989, American Institute of Physics. Reprinted with permission of Oxford University Press, Inc. and American Institute of Physics.]

Surface-emitting lasers have comparable losses and gains to edge-emitting lasers but are much more compact. The need to confine the current to achieve a narrow beam presents problems too. Lateral *p-n* junctions may be used to confine the current. Moreover, the current flowing to the cavity can heat up the mirrors through which it passes, which can lead to severe heating during continuous-wave operation. At present the devices are about 10% efficient in electrical to optical power conversion.

The surface-emitting geometry has many attractions for future systems, including a smaller, circular beam divergence making it easier to couple with optical fibers and integrate with other optoelectronic components.

6.5.15. Low-dimensional lasers

Further improvement in the performance of semiconductor lasers is expected by using higher degrees of quantum confinement – moving from quantum wells (2D) to quantum wires (1D) or quantum dots (0D). One of the driving forces toward such structures is to achieve a "threshold-less" semiconductor laser, i.e. which can reach lasing threshold with minimal electrical current.

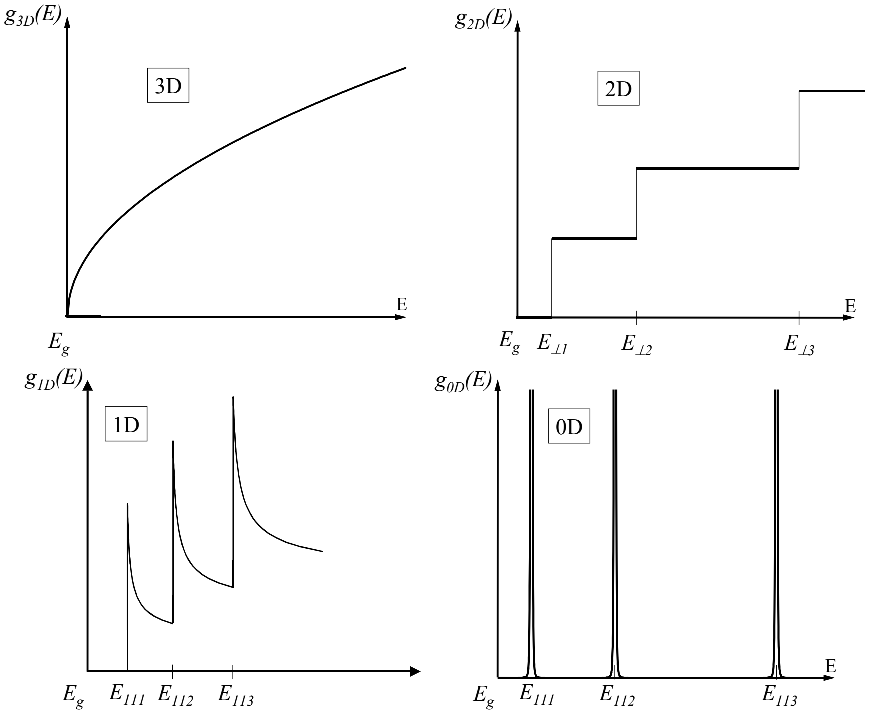


Fig. 6.35. Density of states versus dimensionality: in a bulk (3D) semiconductor crystal, a quantum well (2D), a quantum wire (1D) and a quantum dot (0D).

Fig. 6.35 compares the density of states in bulk crystals (3D), quantum wells (2D), quantum wires (1D), and quantum dots (0D). The very narrow density of states distribution in lower dimensional structures achieves a narrower energy distribution for carriers than in bulk crystals, which results in narrower luminescence spectra, higher differential gain, lower threshold current density and wider modulation bandwidth in lasers using such structures.

Another important feature in quantum dots is the dispersion-less behavior of its electronic states. In other words, the allowed energy levels are independent of the momentum and have a constant energy, which makes it possible to avoid Auger recombination, as long as the positions of the energy levels are positioned adequately. This is illustrated in Fig. 6.36 which compares the Auger recombination process in a bulk semiconductor with that in a quantum dot. If we pretend that the discrete energy levels of the quantum dot lie on a pseudo momentum curve as shown (of course in reality there is no momentum in a quantum dot), we see that, in a quantum dot, the energy and "momentum" conservation cannot be achieved simultaneously during an Auger process. In other words it is difficult for the band to band transition to exactly match an intersubband transition. This ensures that Auger processes in a quantum dots are less likely, which is an important property because a high Auger recombination rate is the major obstacle for the high operating temperature of infrared interband lasers.

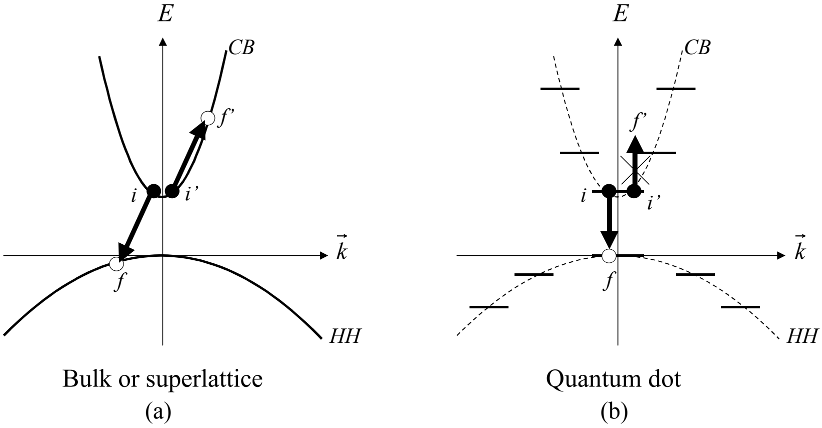


Fig. 6.36. Illustration of the Auger process in a (a) bulk semiconductor or (b) superlattice in comparison with a quantum dot. In the former case, there is a continuum of available states where an Auger electron can be excited into. This makes the conservation of the total "momentum" and energy possible to be achieved in an Auger process. In a quantum dot, there are only discrete energy levels allowed, as a result of the shape of the density of states. Many transitions are therefore forbidden since the" momentum" and energy conservation laws can be satisfied simultaneously only for a few states.

The use of lower dimensional structures has advantages for quantum cascade lasers (QCLs). Indeed, in quantum wires and dots, the LO phonon scattering rates can be considerably lower than in quantum wells. The main reason behind this property is the fact that the scattering rate between two energy bands is proportional to the overlap of the density of states of these bands. Fig. 6.37 shows that such overlap is smaller in a quantum wire and can even be non-existent in a quantum dot. A quantum dot based QCL can therefore have an excellent efficiency through the reduction of phonon scattering rates.

Fabrication of lower dimensional structures is in general extremely difficult. Currently, one of the main targets of atomic engineering research is indeed the fabrication of quantum structures and surfaces with quantum features. The average quantum dot is only 10~50 nm in diameter. Sophisticated growth and/or fabrication techniques are required to produce uniform features in this size regime.

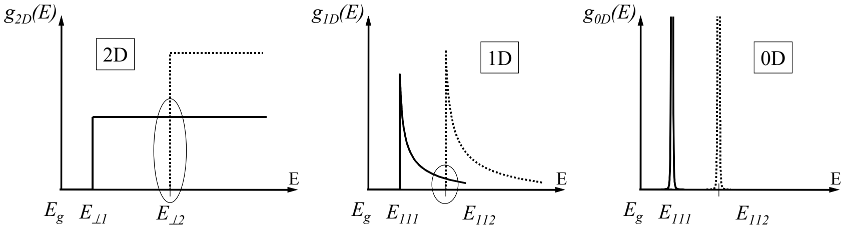


Fig. 6.37. Illustration of the overlap in the density of states for a quantum well (2D), quantum wire (1D), and quantum dot (0D).

6.5.16. Raman lasers

The Raman effect occurs when incident monochromatic light hits the material and a photon is generated with an energy that is different from the incident photon by the energy of a phonon. Specifically, the interaction of an incident photon with the material leads to inelastic scattering where the photon-phonon interactions cause the generation of a photon, which has an energy that is exactly one phonon energy higher or lower than the incident photon. In fact, once there is significant net Raman gain, or photon generation, achieved in a material it is possible to observe lasing action from that material. This effect can be exploited to make lasers from indirect bandgap materials such as silicon, where lasing can be achieved by optically pumping a silicon waveguide. This kind of Raman lasing was first demonstrated in silicon using pulsed lasers in 2004 [Boyraz et al. 2004] and with continuous-wave lasers in 2005 [Rong et al. 2005]. The development of Raman lasing in silicon is considered to be an important milestone in the development of silicon based optoelectronic devices, since silicon is already the most widely used semiconductor in electronic circuits.

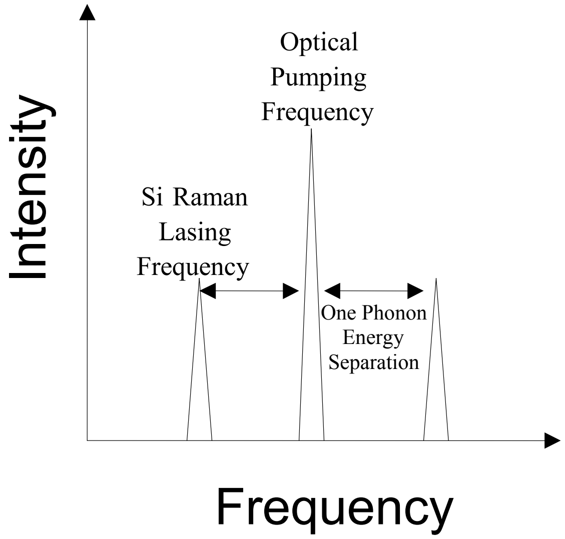

Intensity
Optical Pumping Frequency
Si Raman Lasing Frequency
One Phonon Energy Separation
Frequency

Fig. 6.38. The Raman effect and observed lasing in an optically-pumped silicon waveguide. The separation between the central peak and Raman scattered peaks is the frequency of optical phonons in silicon.

6.6. Summary

In this Chapter, we reviewed the fundamental physical concepts relevant to lasers, including stimulated emission, resonant cavity, waveguide, propagation of an electromagnetic wave in a waveguide and the laser beam divergence, and waveguide design. We introduced the notion of absorption, spontaneous and stimulated emission, the Einstein coefficients, resonant cavity, population inversion and threshold. The example of the ruby laser was used to illustrate these concepts.

The discussion was then focused on semiconductor lasers, which are becoming dominant for numerous modern applications. The concepts of gain, threshold current density, transparency current density, linewidth, external differential quantum efficiency, mirror loss and internal loss were introduced. The different types of semiconductor lasers were then described, including homojunction, single and double heterojunction, and separate confinement and quantum well lasers. The fabrication and packaging technology of semiconductor lasers were then briefly described. Finally, a few specific examples of lasers were presented, including quantum cascade, type II, vertical cavity surface emitting lasers, low-dimensional lasers, and Raman lasers.

References

Casey, H.C., and Panish M.B., *Heterostructure Lasers, parts A & B*, Academic Press, New York, 1978.

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, p. 497, 1995.

Diaz, J., Yi, H.J., Kim, S., Wang, L.J., and Razeghi, M., "High Temperature Reliability of Aluminum-free 980 nm and 808 nm Laser Diodes," *Compound Semiconductors 1995* (*Institute of Physics Conference Series* 145), eds. J.C. Woo and Y.S. Park, Institute of Physics Publishing, Bristol, UK, pp. 1041-1046, 1996.

Diaz, J., Yi, H.J., and Razeghi, M., "Long-term reliability of Al-free InGaAsP/GaAs (λ=808 nm) lasers at high-power high-temperature operation," *Applied Physics Letters* 71, pp. 3042-3044, 1997.

Faist, J., Capasso, F., Sivco, D.L., Hutchinson, A.L., and Cho, A.Y., "Quantum cascade laser," *Science* 264, pp. 553-556, 1994.

Felix, C.L., Meyer, J.R., Vurgaftman, I., Lin, C.H., Murry, S.J., Zhang, D., and Pei, S.S., "High-temperature 4.5 μm type II quantum-well laser with Auger suppression," *IEEE Photonics Technology Letters* 9, pp. 734-736, 1997.

Fuchs, F., Weimer, U., Pletschen, W., Schmitz, J., Ahlswede, E., Walther, M., Wagner, J., and Koidl, P., "High performance InAs/$Ga_{1-x}In_xSb$ superlattice infrared photodiodes," *Applied Physics Letters* 71, pp. 3251-3253, 1997.

Henry, C.H., "Theory of the Linewidth of Semiconductor Lasers", *IEEE Journal of Quantum Electronics* 18, pp. 259-264, 1982.

Holonyak Jr., N. and Bevacqua, S.F., "Coherent (visible) light emission from $Ga(As_{1-x}P_x)$ junctions)," *Applied Physics Letters* 1, pp. 82-83, 1962.

Johnson, J.L., Samoska, L.A., Gossard, A.C., Merz, J., Jack, M.D., Chapman, G.R., Baumgratz, B.A., Kosai, K., and Johnson, S.M., "Electrical and optical properties of infrared photodiodes using the InAs/$Ga_{1-x}In_xSb$ superlattice in heterojunctions with GaSb," *Journal of Applied Physics* 80, pp. 1116-1127, 1996.

Kazarinov, R.F. and Suris, R.A., "Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice," *Soviet Physics Semiconductors* 5, pp. 707-709, 1971.

Kelly, M.J., *Low-Dimensional Semiconductors: Materials, Physics, Technology, Devices*, Oxford University Press, New York, 1995.

Kuznetsov, M., Willner, A.E., Okaminow, I.P., "Frequency-modulation response of tunable 2-segment distributed feedback lasers," *Applied Physics Letters* 55, pp. 1826-1828, 1989.

Lane, B., Wu, A., Stein, A, Diaz, J., and Razeghi, M., "InAsSb InAsP strained-layer superlattice injection lasers operating at 4.0 μm grown by metal-organic chemical vapor deposition," *Applied Physics Letters* 74**,** pp. 3438-3440, 1999.

Lane, B., Tong, S., Diaz, J., Wu, Z., and Razeghi, M., "High power InAsSb/InAsSbP electrical injection laser diodes emitting between 3 and 5 μm," *Material Science and Engineering B* 74**,** pp. 52-55, 2000.

Lin, C.H., Yang, R.Q., Zhang, D., Murry, S.J., Pei, S.S., Allerman, A.A. and Kurtz, S.R., "Type II interband quantum cascade laser at 3.8 μm," *Electronics Letters* 33, pp. 598-599, 1997.

Maiman, T.H., "Stimulated Optical Radiation in Ruby," *Nature* 187, pp. 493-494, 1960.

Mobarhan, K., Razeghi, M., Marquebielle, G., Vassilaki, E., "High-Power 0.98 μm, $Ga_{0.8}In_{0.2}As/GaAs/Ga_{0.51}In_{0.49}P$ Multiple Quantum-Well Laser," *Journal of Applied. Physics* 72, pp. 4447-4448, 1992.

Mohseni, H., Michel, E., Sandven, J., Razeghi, M., Mitchel, W., and Brown, G., "Growth and characterization of InAs/GaSb photoconductors for long wavelength infrared range," *Applied Physics Letters* 71, pp. 1403-1405, 1997.

Boyraz, O. and Jalali, B., "Demonstration of a silicon Raman laser," *Optics Express* 12, pp. 5269-5273, 2004.

Razeghi, M., Hirtz, P., Blondeau, R., and Duchemin, J.P., "Aging Test of MOCVD Shallow Proton Stripe GaInAsP-InP, DH Laser Diode Emitting at 1.5 μm," *Electronics Letters* 19, p. 481, 1983a.

Razeghi, M., Hersee, S., Blondeau, R., Hirtz, P., and Duchemin, J.P., "Very Low Threshold GaInAsP/InP DH Lasers Grown by MOCVD," *Electronics Letters* 19, p. 336, 1983b.

Razeghi, M., in *Lightwave Technology for Communication*, ed. W.T. Tsang, Academic Press, New York, 1985a.

Razeghi, M., Blondeau, R., Boulay, J.C., de Cremoux, B., and Duchemin, J.P., "LP-MOCVD growth and CW operation of high quality SLM and DFB semiconductor $Ga_xIn_{1-x}As_yP_{1-y}$–InP lasers," in *GaAs and Related Compounds 1984* (*Institute of Physics Conference Series* 74), UK Adam Hilger, Bristol, UK, p. 451, 1985b.

Razeghi, M., Blondeau, R., Krakowski, M., Bouley, J.C., Papuchon, M., de Cremoux, B., and Duchemin, J.P. "Low-Threshold Distributed Feedback Lasers Fabricated on Material Grown Completely by LP-MOCVD," *IEEE Journal of Quantum Electronics* QE-21, pp. 507-511, 1985c.

Razeghi, M., "CW Phase-Locked Array GaInAsP-InP High Power Semiconductor Laser Grown by Low- Pressure Metalorganic Chemical Vapor Deposition," *Applied Physics Letters* 50, p. 230, 1987.

Razeghi, M., "High-power laser diodes based on InGaAsP alloys," *Nature* 369, pp. 631-633, 1994.

Razeghi, M., "High Power InAsSb/ InAsSbP Laser Diodes Emitting in the 3-5 μm Range," in *1998 Army Research Office Highlights*, Physical Sciences Directorate, 1998.

Razeghi, M., Wu, D., Lane, B., Rybaltowski, A., Stein, A., Diaz, J., and Yi, H., "Recent achievement in MIR high power injection laser diodes (λ=3 to 5 μm)," *LEOS Newsletter* 13, pp. 7-10, 1999.

Rong, H., Jones, R., Liu, A., Cohen, O., Hak, D., Fang, A., and Paniccia, M., "A continuous-wave silicon Raman laser," *Nature* 433, pp. 725-728, 2005.

Sai-Halasz, G.A., Tsu, R., and Esaki, L., "A new semiconductor superlattice," *Applied Physics Letters* 30, pp. 651-653, 1977.

Sai-Halasz, G.A., Chang, L.L., Welter, J.M., Chang, C.A., and Esaki, L., "Optical absorption of $In_{1-x}Ga_xAs-GaSb_{1-y}As_y$ superlattices," *Solid State Communications* 27, pp. 935-937, 1978a.

Sai-Halasz, G.A., Esaki, L., and Harrison, W.A., "InAs-GaSb superlattice energy structure and its semiconductor-semimetal transition," *Physical Review B* 18, pp. 2812-2818, 1978b.

Schawlow, A.L. and Townes, C.H., “Infrared and optical masers,” *The Physical Review* 112, pp. 1940-1949, 1958.

Smith, D.L., and Mailhiot, C., “Proposal for strained type II superlattice infrared detectors,” *Journal of Applied Physics* 62, pp. 2545-2548, 1987.

Wu, D., Lane, B., Mosheni, H., Diaz, J., and Razeghi, M., “High power asymmetrical InAsSb/ InAsSbP/ AlAsSb double heterostructure lasers emitting at 3.4 μm,” *Applied Physics Letters* 74**,** pp. 1194-1196, 1999.

Xie, H., Wang, W.I., and Meyer, J.R., “Infrared electroabsorption modulation at normal incidence in asymmetrically stepped AlSb/InAs/GaSb/AlSb quantum wells,” *Journal of Applied Physics* 76, pp. 92-96, 1994.

Yang, B.H., Zhang, D., Yang, R.Q., Lin, C.H., Murry, S.J., and Pei, S.S., “Mid-infrared interband cascade lasers with quantum efficiencies > 200%,” *Applied Physics Letters* 72, pp. 2220-2222, 1998.

Yi, H., Diaz, J., Wang, L.J., Kim, S., Williams, R., Erdtmann, M., He, X., and Razeghi, M., “Optimized structure for InGaAsP/GaAs 808 nm high power lasers,” *Applied Physics Letters* 66, pp. 3251-3253, 1995.

Youngdale, E.R., Meyer, J.R., Hoffman, C.A., Bartoli, F.J., Grein, C.H., Young, P.M., Ehrenreich, H., Miles, R.H., and Chow, D.H., “Auger lifetime enhancement in InAs-$Ga_{1-x}In_xSb$ superlattices,” *Applied Physics Letters* 64, pp. 3160-3162, 1994.

Further reading

Agrawal, G. and Dutta, N., *Semiconductor Lasers*, Van Nostrand Reinhold, New York, 1993.

Felix, C.L., Meyer, J.R., Vurgaftman I., Lin, C.H., Murry, S.J., Zhang, D., and Pei, S.S., “High-temperature 4.5 μm Type-II quantum-well laser with Auger suppression,” *IEEE Photonics Technology Letters* 9, pp. 734-736, 1997.

Iga, K., *Fundamentals of Laser Optics*, Plenum Press, New York, 1994.

Johnson, J.L., Samoska, L.A., Gossard, A.C., Merz, J., Jack, M.D., Chapman, G.R., Baumgratz, B.A., Kosai, K., and Johnson, S.M., “Electrical and optical properties of infrared photodiodes using the InAs/$Ga_{1-x}In_xSb$ superlattice in heterojunctions with GaSb,” *Journal of Applied Physics* 80, pp. 1116-1127, 1996.

Kim, S. and Razeghi, M., "Recent advances in quantum dot optoelectronic devices and future trends," in *Handbook of Advanced Electronic and Photonic Materials and Devices*, ed. H.S. Nalwa, Academic Press, London, pp. 133-154, 2001.

Lin, C.H., Yang, R.Q., Zhang, D., Murry, S.J., Pei, S.S., Allerman, A.A. and Kurtz, S.R., “Type-II interband quantum cascade laser at 3.8 μm,” *Electronics Letters* 33, pp. 598-599, 1997.

Mohseni, H., Michel, E., Sandven, J., Razeghi, M., Mitchel, W., and Brown, G., “Growth and characterization of InAs/GaSb photoconductors for long wavelength infrared range,” *Applied Physics Letters* 71, pp. 1403-1405, 1997.

O'shea, D., *Introduction to Lasers and Their Applications*, Addison-Wesley, Reading, MA, 1978.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for Photonic and Electronic Device Applications*, Institute of Physics, Bristol, UK, pp. 21-29, 1995.

Razeghi, M., "Optoelectronic Devices Based on III-V Compound Semiconductors Which Have Made a Major Scientific and Technological Impact in the Past 20 Years," *IEEE Journal of Selected Topics in Quantum Electronics*, 2000.

Razeghi, M., Wu, D., Lane, B., Rybaltowski, A., Stein, A., Diaz, J., and Yi, H., "Recent achievements in MIR high power injection laser diodes (λ = 3 to 5 µm)," *LEOS Newsletter* 13, pp. 7-10, 1999.

Razeghi, M., "Kinetics of Quantum States in Quantum Cascade Lasers: Device Design Principles and Fabrication," *Microelectronics Journal* 30, pp. 1019-1029, 1999.

Scherer, A., Jewell, J., Lee, Y.H., Harbison, J., and Florez, L.T., "Fabrication of microlasers and microresonator optical switches," *Applied Physics Letters* 55, pp. 2724-2726, 1989.

Siegman, A.E., *Lasers*, University Science Book, Mill Valley, Calif., 1986.

Silfvast, W.T., *Laser Fundamentals*, Cambridge University Press, New York, 1996.

Streetman, B.G., *Solid States Electronic Devices,* Prentice-Hall, Englewood Cliffs, NJ, 1990.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Problems

1. In a space mission to Mars, it is needed to equip the research robot with a single mode laser to make sample absorbance/transmittance analysis to identify the unknown gases. Assume that all types of lasers (e.g. gas, solid state, liquid, semiconductor lasers) specified in the Chapter are offering wavelength ranges and power levels adequate for this application. Which one would you choose for this application? Specify which properties of this type of laser make it more suitable?

2. Using the definition of the coefficients B_{12} and B_{21} given in sub-section 6.3.1, show that $B_{12} = B_{21}$. Take the photon field $\rho(E_{21})$ as $N(E_{21})n_{ph}$ where $N(E_{21})$ is the density of energy states for the photon field and n_{ph} is the average number of photons defined as $n_{ph} = \frac{1}{e^{E_{21}/k_b T} - 1}$.

3. A laser diode has a gain peak between 4.45–4.65 µm, and we want to have only a single mode at 4.5 µm as output of this laser diode. What should be the length of the laser so that only a single mode is allowed

within the laser? Take the refractive index within the active region as 3.2. (This type of laser is called short cavity laser).

4. Find the spacing (free spectral width) of wavelengths between adjacent modes in a Fabry Perot cavity. Compare this with the mode spacing in terms of frequency. Which one would you prefer to use if you are comparing free spectral widths of two cavities designed for different wavelength regimes?

5. *Duality principle.*
Once the solutions for E field are known, solutions for the H field can be found easily by using duality principle. Change the source free Maxwell's equations (Eq. (6.17)) by using the following substitutions: $E \rightarrow H$, $H \rightarrow -E$, $\varepsilon \rightarrow \mu$, $\mu \rightarrow \varepsilon$. Does anything change?

6. Consider an edge-emitting laser with identical facets. Find the mirror loss and threshold gain if $\overline{n}_{active} = 3.2$, $\alpha_I = 5\ \mathrm{cm}^{-1}$ and $L = 1$ mm. Now consider that one of the facets is coated with a dielectric/metal layer such that the reflection coefficient is $R = 0.98$, find threshold gain and mirror loss for this case and compare with the previous result. Which one do you think will have a smaller threshold current and higher power when the output from the facet with the same R is measured? Why?

7. Plot the external quantum efficiency η_e^{-1} vs. L when L is varying from 0 to 3 mm, given the values $\alpha_I = 5\ \mathrm{cm}^{-1}$ and $\overline{n}_{active} = 3.2$ and $\eta_e^{-1}(L) = 2.0327$, 2.8154 and 3.5981 for $L = 1$, 2 and 3 mm respectively. Find the internal quantum efficiency η_i^{-1}.

8. For a semiconductor laser diode, it is possible to make an edge-emitting laser be surface emitting at the same time by using second order corrugations on the waveguide and laterally inducing current so the surface of the ridge is not coated with metal. Consider the far field characteristics of the edge emission given the aperture sizes around 2 μm by 40 μm for the edge. Compare to a surface emission aperture with the same width, but a length of 35 μm. Which one do you expect to have less divergence and what is the beam shape in each case?

9. Quantum cascade lasers do not suffer Auger processes due to the fact that intersubband interactions involve electrons only, and in the cladding layers the majority carriers are electrons. What is the dominant loss mechanism in this type of laser?

7. Quantum Cascade Lasers

7.1. Introduction
7.2. Basic operation principles
  7.2.1. Intersubband transitions
  7.2.2. Cascading
  7.2.3. Rate equation
  7.2.4. Polar optical phonon resonance
7.3. The components of a quantum cascade laser
  7.3.1. Core heterostructure
  7.3.2. Laser waveguide
7.4. Making a quantum cascade laser
  7.4.1. Epitaxial growth and material characterization
  7.4.2. Processing and packaging
7.5. Device performances
  7.5.1. Power-current-voltage characteristics
  7.5.2. Temperature dependent characteristics
  7.5.3. Wall plug efficiency
  7.5.3. Spectra and far field
7.6. Wall plug efficiency optimizations
  7.6.1. Electrical contact resistance
  7.6.2. Waveguide geometry
  7.6.3. Bonding method
7.7. Power scaling
7.8. Photonic crystal distributed feedback quantum cascade lasers
  7.8.1. Pattern design
  7.8.2. Coupling coefficients
  7.8.3. Testing results
7.9. Quantum cascade lasers at different wavelengths
  7.9.1. Short wavelength quantum cascade lasers (< 4 μm)
  7.9.2. Mid wavelength quantum cascade lasers (4–9 μm)
  7.9.3. Long wavelength quantum cascade lasers (>9 μm)
7.10. Summary

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_7,
© Springer Science+Business Media, LLC 2010

7.1. Introduction

As a result of mass production, millions of semiconductor diode lasers are manufactured each month and appear in products ranging from telecommunications transmitters to DVD players and laser pointers. For traditional laser diodes, the applications are often dictated by what part of the electromagnetic spectrum is accessible. For example, telecommunications lasers operate in a region of the infrared where silica optical fiber has minimum dispersion or transmission loss. Laser-based displays, on the other hand, require red, green, and blue lasers to make a visible image. A biological fluorescence system will often require an ultraviolet source to function correctly. Though most people are unaware, a large part of the electromagnetic spectrum is still not fully utilized commercially due to the lack of a proper laser source. This includes the bulk of the “infrared” region. Though the scientific community has been exploring it for some time, the systems used for research are usually too bulky, too expensive, and too hard to understand to become large scale commercial products.

Perhaps the most important feature, and one of the main driving forces of quantum cascade laser (QCL) development until recently, is the ability of all molecules to absorb in the infrared. The atoms in a molecule can bend, stretch, and rotate with respect to one another, and these excitations, to a large part, are optically active. Most molecules, from simple to moderately complex, have a characteristic absorption spectrum in the 3–14 μm wavelength range from which they can be uniquely identified and quantified in real time. In larger concentrations, infrared spectroscopy has been used for this purpose by industry for many years. The benefit of the optical technique, as opposed to chemical sensors or chromatography, is that the detection mechanism requires minimal sample pre-treatment, is re-usable, and is very fast.

For hazardous chemicals, such as explosives or nerve agents, we need to be aware when even the smallest amount is present. Common industrial spectroscopy systems do not have the sensitivity to detect trace amounts, which is why laser systems are necessary. Laser systems allow for both direct, long-path-length absorption measurements as well as indirect monitoring of chemicals through high power excitation (photoacoustic and photothermal detection). These systems can be built in many ways, but can quantify specific chemicals down to part-per-billion concentrations.

Besides simple spectroscopy, this direct absorption by specific molecules also lends itself to some other potentially useful medical technologies. Breath analysis, for example, has already been used to monitor health by checking for abnormal cell metabolism byproducts. In the future, it may also be possible to target or cauterize specific types of cells by their chemical or protein content for selective surgery.

The second appeal of this wavelength range is related to atmospheric transmission. While the above applications were based on the presence of absorbing species, this application deals with their absence. There are a couple of broad wavelength ranges (3–5 and 8–12 μm) that have minimal absorption by typical atmospheric gases. In the near-infrared, there has been a lot of research into free-space optical links to complete the "last-mile" connectivity problem for high speed internet. These systems have good bandwidth and are cheaper to install than fiber optic cable. One downside to these systems is their sensitivity to weather. Scattering by rain and fog severely attenuates the signal which reduces the bandwidth, distance, and/or availability of the data link. At longer wavelengths, it has been proven that the scattering by fog is dramatically reduced. While not completely eliminated, an inexpensive midwave or longwave laser source could replace the near-infrared laser and lead to better range and availability for these systems. At the same time, if the power can be increased without adding significantly to cost or causing eye safety issues, the link distance will be proportionally longer as well.

The original idea for the QCL was proposed by Kazarinov and Suris [Kazarinov et al. 1971]. Since the first experimental demonstration [Faist et al. 1994], QCLs have been experiencing rapid development. QCLs with room temperature continuous wave (CW) operation capable of delivering more than 100 mW optical power have been demonstrated at a number of wavelengths between 3.8 μm [Yu et al. 2006] and 11.5 μm [Slivken et al. 2008]. At a particular wavelength near 4.7 μm, significant effort was devoted to improve the output power [Bai et al. 2008a] and the wall plug efficiency [Bai et al. 2008b] of this kind of device.

7.2. Basic operation principles

The uniqueness of a QCL is that the laser transition takes place between two intersubband states, whose energy difference can be engineered to produce lasers with different wavelengths using the same material system. In addition, the electrons are recycled by a cascading design, which allows for multiple photons emitted by a single electron. Aided by a simple rate equation model, the population inversion can be achieved by engineering the lifetime of the lower laser level significantly shorter than that of the upper laser level. This can be obtained by making use of the polar optical phonon resonance.

7.2.1. Intersubband transitions

An illustration of the differences between interband and intersubband designs is shown in Fig. 7.1. The interband structure utilizes that natural band gap for electron transitions. As such the material composition controls mainly the emission energy. In an intersubband emitter, a new "band gap" between quantum well subbands is created for this purpose. Since the spacing between subbands is controlled by the quantum well thickness, the emission energy or wavelength can be tuned over a very wide range without changing the material composition. This has two main benefits. First, as the emission is material independent, we can use the most mature and robust material technology available as the basis of our device (no exotic materials). Second, because the curvature of the bands is the same, the efficiency of photon emission has a much weaker dependence on wavelength and temperature, which allows easier access to infrared lasers at room temperature.

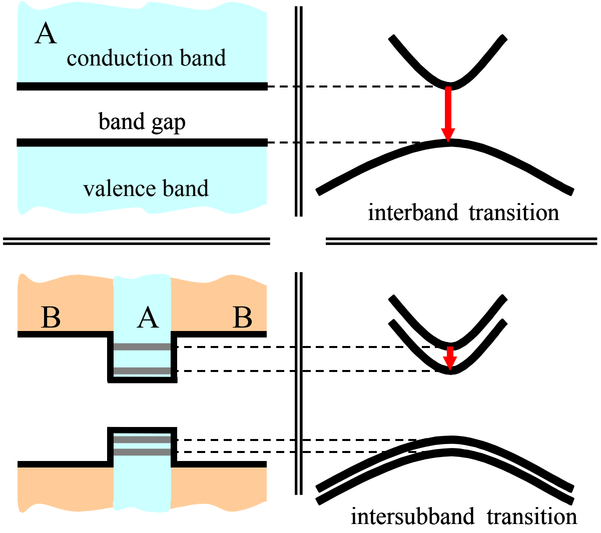


Fig. 7.1.Illustration of the differences between interband and intersubband transitions. Conventional diode lasers are based on the interband transition. Whereas the quantum cascade lasers are based on the intersubband transition. [Reproduced with permission from Optics and Photonics News July 2008, M. Razeghi, "The Quantum Cascade Laser, a Versatile and Powerful Tool," pg. 45, Copyright 2008, Optics Society of America (OSA).)]

Of course, the structure of a practical QCL is much more complicated than that depicted in Fig. 7.1. But the underling physics principle is similar. Instead of only three layers with one quantum well (material A) switched by two quantum barriers (material B), the heterostructure for one QCL stage contains tens of layers, with the thickness of each layers engineered in such a way that the desired subbands are formed. This includes not only the eigen energies and the shape of the eigen-functions for the two main subbands responsible for the laser transition, but also those for all other subbands, which are related to how well the laser functions. For a QCL structure containing tens of layers, numerical Schrödinger solvers are usually developed for fast calculations.

7.2.2. Cascading

The QCL is a subset of a general intersubband laser which specifies not only the light emission mechanism, but electron transport and power scaling as well. While the earliest intersubband emitters were optically pumped, the QCL is electrically pumped (e.g. with a battery), similar to standard laser diodes. The essence is to engineer the electron flow through a series of subbands and minibands (groups of subbands) which allow for a buildup of electrons within the upper laser subband, and fast extraction from the lower subbands. This is accomplished by a special injector region which blocks electron escape out the upper subband, yet couples strongly to the lower subband. In addition, this injector region, as the name implies, also serves to recycle extracted electrons by re-injecting them into the upper laser subband of a subsequent emitting stage. An illustration of this process is shown in Fig. 7.2. This quantum-based transport combined with the cascaded emitter design is responsible for the name of the device.

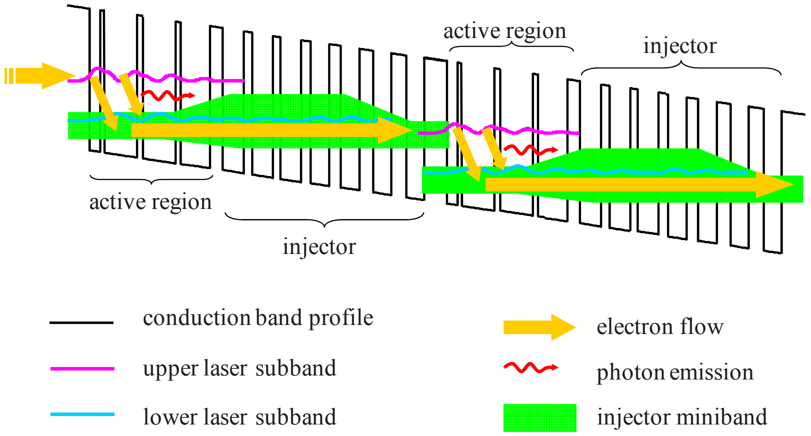


Fig. 7.2.Two emitting stages of the quantum cascade laser. This illustrates the role of the injector region and the cascaded nature of the photon emission, whereas a single electron can emit multiple photons[Reproduced with permission from Optics and Photonics News July 2008, M. Razeghi, "The Quantum Cascade Laser, a Versatile and Powerful Tool," pg. 46, Copyright 2008, Optics Society of America (OSA).)]

The general sloped conduction band profile in Fig. 7.2 is to simulate the consequence of the external field applied to the device during operation. The injector miniband consists of many strongly coupled subbands, similar to that of a superlattice. But, in this case it is a chirped superlattice, meaning the miniband is only formed for a range of applied external field. When the external field is too low or too high, the coupling among those subbands is not strong enough to form a miniband and they need to be treated as individual subbands. Note that the plot of Fig. 7.2 is in real space and the energy dispersion as a function of the in-plane momentum is not shown. In fact, the simulation assumes zero in-plane momentum and the subband nature, i.e., a range of energy, of each state is implied. As such, the subbands in a QCL system are also referred to as levels as a means to simplify the laser simulation.

7.2.3. Rate equation

The full analysis for the QCL system is far too complex to be treated here. Even with the knowledge of eigen-energies and eigen-functions for all the subbands, populating these subbands is not trivial as electrons are not in thermal equilibrium among different subbands in an operating QCL system. Instead, distribution of electrons among subbands is governed by a steady-state condition, in which the rate of electrons scattered into a subband equals the rate scattered out of it. In a layered semiconductor system, there are numerous types of interactions responsible for such scattering mechanisms. Thus one needs to know the importance of all these scattering mechanisms, including the polar optical phonon scattering, electron-electron scattering, interface roughness scattering, impurity scattering, and alloy scattering. Besides the large number of subbands, this problem can be difficult to solve because some of these scattering processes depend on the unknown population of the related subbands, e.g., the electron-electron scattering. As a result, a self-consistent solution needs to be reached by iteration, which dramatically increases the burden of the numerical simulation.

In order to approximate the carrier dynamics for some of the most important subbands, a simplified rate equation model is developed, yet is capable of revealing some of the most important device performance parameters, such as the differential gain and the internal quantum efficiency.

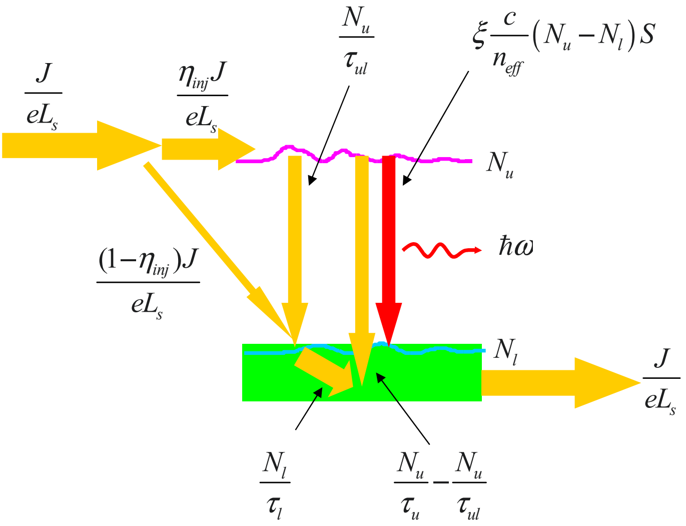

$\frac{J}{eL_s}$
$\frac{\eta_{inj}J}{eL_s}$
$\frac{N_u}{\tau_{ul}}$
$\xi\frac{c}{n_{eff}}(N_u-N_l)S$
N_u
$\hbar\omega$
$\frac{(1-\eta_{inj})J}{eL_s}$
N_l
$\frac{J}{eL_s}$
$\frac{N_l}{\tau_l}$
$\frac{N_u}{\tau_u}-\frac{N_u}{\tau_{ul}}$

Fig. 7.3. Simplified rate equation model for the active region of the QCL structure.

Let us focus on the active region of Fig. 7.2, which is enlarged in Fig. 7.3. The total current density flowing into the active region is denoted by J, of which a major part is tunnel injected into the upper laser level with an injection efficiency of η_{inj}. The rest of it is injected into the lower laser level. The population of the upper laser level and the lower laser level is denoted by N_u and N_l, respectively. τ_u, τ_l, and τ_{ul} are lifetimes related to electrons scattered out of the upper laser level, out of the lower laser level and from the upper laser level into the lower laser level, respectively. Note that the scattering process responsible for τ_{ul} is part of those responsible for τ_u, since there are many levels located within the injector miniband. The intersubband transition induced by the stimulated emission is represented by a term containing the photon density S. This term is also proportional to the population inversion $N_u - N_l$ and the gain coefficient ξ.

The rate equation below threshold, i.e., without the stimulated emission term, is constructed following the simple fact that the changing rate of the population for a certain level equals the difference of the scattering rate into this level and that out of this level. As such we have the following rate equations for the upper laser level and the lower laser level:

Eq. (7.1) $$\frac{dN_u}{dt}=\frac{J\eta_{inj}}{eL_s}-\frac{N_u}{\tau_u}$$

Eq. (7.2) $$\frac{dN_l}{dt}=\frac{J(1-\eta_{inj})}{eL_s}+\frac{N_u}{\tau_{ul}}-\frac{N_l}{\tau_l}$$

where e and L_s are electron charge and the length of one QCL stage, respectively.

Eq. (7.1) and Eq. (7.2) describe the dynamic behavior of N_u and N_l, but here we are only interested in the steady state solution, i.e., N_u and N_l do not change with time. As such we will set $\frac{dN_u}{dt}$ and $\frac{dN_l}{dt}$ to zero, resulting the following set of equations:

Eq. (7.3) $$\begin{cases} 0=\dfrac{J\eta_{inj}}{eL_s}-\dfrac{N_u}{\tau_u} \\ 0=\dfrac{J(1-\eta_{inj})}{eL_s}+\dfrac{N_u}{\tau_{ul}}-\dfrac{N_l}{\tau_l} \end{cases}$$

Assuming all parameters are known except N_u and N_l, we can solve Eq. (7.3) to have:

Eq. (7.4) $$N_u=\frac{\eta_{inj}J}{eL_s}\tau_u$$

Eq. (7.5) $$N_l=\frac{\eta_{inj}J}{eL_s}\left[\tau_u-\left(\frac{1}{\eta_{inj}-\eta_i}-\frac{1}{\eta_{inj}}\right)\tau_l\right]$$

with the internal quantum efficiency η_i defined as:

Eq. (7.6) $$\eta_i=\eta_{inj}-\frac{\tau_l}{\tau_u(1-\frac{\tau_l}{\tau_{ul}})+\tau_l}$$

Eq. (7.5) and Eq. (7.6) give the population inversion:

Eq. (7.7) $$\Delta N = N_u - N_l = \frac{\eta_{inj} J}{eL_s}\left(\frac{1}{\eta_{inj}-\eta_i} - \frac{1}{\eta_{inj}}\right)\tau_l$$

As we can see for $\eta_i > 0$, ΔN increases with J . When J reaches a certain value, i.e., the threshold current density, the total optical loss of the system α_{total} will be completely compensated by the optical gain of the system:

Eq. (7.8) $$g = \xi\left(N_u - N_l\right)\Gamma$$

where Γ is the optical confinement factor within the gain medium.

Combining Eq. (7.8) and Eq. (7.7), we get the differential gain G_d:

Eq. (7.9) $$G_d = \frac{dg}{dJ} = \frac{\xi\Gamma}{eL_s}\left\{\left[\tau_u\left(1-\frac{\tau_l}{\tau_{ul}}\right)+\tau_l\right]\eta_{inj} - \tau_l\right\}$$

For above threshold operation, we need to add the stimulated emission term into Eq. (7.1) and Eq. (7.2), resulting:

Eq. (7.10) $$\frac{dN_u}{dt} = \frac{J\eta_{inj}}{eL_s} - \frac{N_u}{\tau_u} - \xi\frac{c}{n_{eff}}\left(N_u - N_l\right)S$$

Eq. (7.11) $$\frac{dN_l}{dt} = \frac{J(1-\eta_{inj})}{eL_s} + \frac{N_u}{\tau_{ul}} - \frac{N_l}{\tau_l} + \xi\frac{c}{n_{eff}}\left(N_u - N_l\right)S$$

In addition, a similar rate equation for the photon density can be constructed considering the balance of photon generation and elimination:

Eq. (7.12) $$\frac{dS}{dt} = S\frac{c}{n_{eff}}\left[\xi\left(N_u - N_l\right)\Gamma - \alpha_{total}\right]$$

The set of equations for the steady state reads:

Eq. (7.13)
$$\begin{cases} 0 = \dfrac{J\eta_{inj}}{eL_s} - \dfrac{N_u}{\tau_u} - \xi \dfrac{c}{n_{eff}} (N_u - N_l) S \\ 0 = \dfrac{J(1-\eta_{inj})}{eL_s} + \dfrac{N_u}{\tau_{ul}} - \dfrac{N_l}{\tau_l} + \xi \dfrac{c}{n_{eff}} (N_u - N_l) S \\ 0 = S \dfrac{c}{n_{eff}} \left[\xi (N_u - N_l) \Gamma - \alpha_{total} \right] \end{cases}$$

Again, we treat all parameters as known except N_u, N_l and S to solve Eq. (7.13). The solution is:

Eq. (7.14)
$$N_u = \frac{\eta_{inj} J_{th}}{eL_s} \tau_u + \frac{J - J_{th}}{eL_s} \tau_u \left(\eta_{inj} - \eta_i \right)$$

Eq. (7.15)
$$N_l = \frac{\eta_{inj} J_{th}}{eL_s} \left[\tau_u - \left(\frac{1}{\eta_{inj} - \eta_i} - \frac{1}{\eta_{inj}} \right) \tau_l \right] + \frac{J - J_{th}}{eL_s} \tau_u \left(\eta_{inj} - \eta_i \right)$$

Eq. (7.16)
$$S = \frac{\Gamma \eta_i}{eL_s \dfrac{c}{n_{eff}} \alpha_{total}} \left(J - J_{th} \right)$$

with the threshold current density J_{th} defined as:

Eq. (7.17)
$$J_{th} = \frac{eL_s \alpha_{total}}{\xi \Gamma} \Bigg/ \left\{ \left[\tau_u \left(1 - \frac{\tau_l}{\tau_{ul}} \right) + \tau_l \right] \eta_{inj} - \tau_l \right\}$$

The population inversion above threshold can be obtained from Eq. (7.14) and Eq. (7.15), which is:

Eq. (7.18)
$$\Delta N = N_u - N_l = \frac{\eta_{inj} J_{th}}{eL_s} \left(\frac{1}{\eta_{inj} - \eta_i} - \frac{1}{\eta_{inj}} \right) \tau_l$$

As expected, ΔN is independent of J, and clamped at the value when $J = J_{th}$, if we compare Eq. (7.18) with Eq. (7.7).

In addition, the upper laser level population at threshold can be obtained from Eq. (7.4) as:

Eq. (7.19) $$N_u^{th} = \frac{\eta_{inj} J_{th}}{eL_s} \tau_u$$

Combining Eq. (7.19) and Eq. (7.14), we have:

Eq. (7.20) $$N_u - N_u^{th} = \frac{J - J_{th}}{eL_s} \tau_u \left(\eta_{inj} - \eta_i \right)$$

If we compare Eq. (7.20) with Eq. (7.4), a very important conclusion can be drawn, i.e., the lifetime of the upper laser level is reduced from τ_u (below threshold) to $\tau_u \left(\eta_{inj} - \eta_i \right)$ (above threshold) due to the presence of the stimulated emission.

The output power P is proportional to the photon density S and the mirror loss α_m by:

Eq. (7.21) $$P = \hbar\omega \left(\frac{NAL_s}{\Gamma} \right) \left(\frac{c}{n_{eff}} \right) S\alpha_m$$

with N and A denoting the number of QCL stages and the area of the current cross section, i.e., the product of the cavity length and the ridge width.

Inserting Eq. (7.16) into Eq. (7.21) and writing the total loss α_{total} as the sum of the mirror loss α_m and the waveguide loss α_w, we have the expression for the slope efficiency:

Eq. (7.22) $$\eta_s = \frac{dP}{dI} = \frac{N\hbar\omega}{e} \eta_i \frac{\alpha_m}{\alpha_m + \alpha_w}$$

Eq. (7.22) finally reveals the physical meaning of the internal quantum efficiency (Eq. (7.6)), which is the portion of electrons that contribute to the photon emission out of all the electrons that contribute to the transport. It can also be interpreted as the number of photons emitted per injected electron per emitting stage.

At this point, a general picture for the upper and lower laser level population as a function the current density can be described as the following. Below threshold, the population of both the upper laser level and the lower laser level increases as the current density increases (Eq. (7.4) and Eq. (7.5)). The population of the upper laser level increases faster than that of the lower laser level, resulting a growing of the population inversion as a function of the current density (Eq. (7.7)). Since the gain is proportional to the population inversion (Eq. (7.8)), at some point the device will exhibit enough gain to compensate for the total optical loss (absorption and scattering) of the system and the device reaches the threshold condition. Above threshold, the populations of both the upper laser level and the lower laser level continue to increase as a function of the current density (Eq. (7.14) and Eq. (7.15)), but at the same rate. This gives a clamped population inversion. (Eq. (7.18)) Due to the presence of the stimulated emission above threshold, the lifetime of the upper laser level is reduced by a factor related to the injection efficiency and the internal quantum efficiency.

It needs to be made clear that this simplified rate equation model has limitations. First of all, this model assumes a “perfect” injector, i.e., without backfilling the lower laser level and free of transport delay. But this is not the case for a real device operating at room temperature. Thermal backfilling and injector transport delay may play an important role. Secondly, there is no carrier conservation in this model. The population of both the upper and lower laser level can keep increasing without causing depletion somewhere else in the structure as if there is a big reservoir. For a real device, a saturation effect is expected since the doping level of a QCL is usually not very high and the subband alignment is only maintained for a limited electric field range. Thirdly, the injection efficiency and lifetimes are assumed to be constant as a function of current density. This is a good approximation only when the device is operated in the vicinity of the threshold. For substantially higher current density above the threshold, these parameters are expected to change somewhat. Lastly, the above model also fails at zero current density. It gives zero population for both the upper laser level and the lower laser level at zero current density, which is obviously not the case for a real device. In fact, the lower laser level will be heavily populated by thermal excitation at zero current density.

In spite of all these limitations, this simple rate equation model is a good starter for students to grasp the basic operation principles off a QCL.

7.2.4. Polar optical phonon resonance

One of the most important results of the rate equation model is the expression for the internal quantum efficiency (Eq. (7.6)), which fundamentally determines how well the device operates. The least requirement for lasing is a positive internal quantum efficiency as shown by the expression for the population inversion (Eq. (7.7)). A functional QCL also requires an internal quantum efficiency above a certain value such that the population inversion can be big enough to generate sufficient gain to compensate for optical losses. To make a better QCL, we want the internal quantum efficiency as close to unity as possible.

Let us examine Eq. (7.6) closely. First of all, we want the injection efficiency as close to unity as possible. This can be achieved by designing the coupling of the injector to the active region of the next stage in such a way that it strongly couples to the upper laser level, whereas weakly coupled to the lower laser level, which is illustrated in Fig. 7.4. The first thin well in the active region right next to the injection barrier serves as an energy filter, which enhances (suppresses) the electron transmission from the injector to the upper (lower) laser level by making the upper laser level penetrate more into the injection barrier and separate the lower laser level more from the injection barrier.

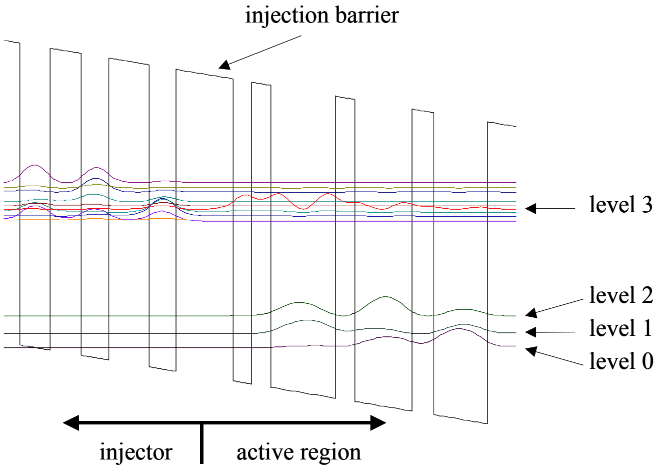


Fig. 7.4. A portion of the QCL structure in real space showing the coupling between the injector and the active region. Level 3 is the upper laser level and couples strongly to the injector miniband. Level 2 is the lower laser level and couples weakly to the injector miniband. The optical phonon resonance scheme is achieved between level 2 and level 1.

From the term containing the lifetimes in Eq. (7.6), it follows that we want a big τ_u, a big τ_{ul} and a small τ_l. These lifetimes are the consequence of many kinds of scattering events taking place in the QCL system, among which the polar optical phonon (POP) scattering is the most important scattering mechanism within the active region. The polar optical phonon scattering needs to fulfill both the energy conservation and the momentum conservation. The energy difference between the initial and final states equals to the POP energy ε_{ph}. The momentum of the participating POP equals to the momentum difference Δp between the initial and final states. As depicted in Fig. 7.5, if we purposely make the energy separation between level 2 and level 1 close to ε_{ph}, we can have a very small Δp_l, which goes to zero at POP resonance, i.e., the energy separation between level 2 and level 1 equals to ε_{ph}. The theory for the POP scattering [Smet et al. 1996] suggests that a smaller momentum difference Δp gives a smaller lifetime τ. Therefore, we can make the lifetime of the lower laser level much smaller than that of the upper laser level using a POP resonance scheme.

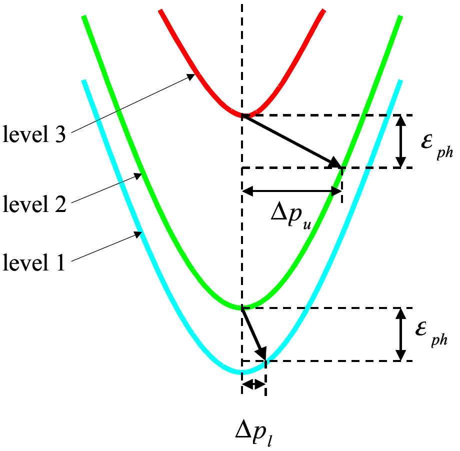


Fig. 7.5. Illustration of the active region states in the momentum space. Level 3 is the upper laser level and level 2 is the lower laser level.

According to Eq. (7.6), only the relative magnitude of those lifetimes are important for the internal quantum efficiency. Although all of them are on the order of a picosecond, which is more than 6 orders of magnitude faster than the spontaneous emission, lasing is not precluded for QCLs as long as we can make the lower laser level lifetime significantly smaller than the upper laser level lifetime. The POP resonance scheme has been proven to be one of the best choices for efficient QCL designs.

7.3. The components of a quantum cascade laser

Like other semiconductor lasers, a quantum cascade laser consists of the core heterostructure and the laser waveguide.

7.3.1. Core heterostructure

Unlike some traditional semiconductor devices, the QCL is truly a "quantum" device. The individual layer thicknesses, as shown in Fig. 7.12, range only 1–4 nm. With tens of layers per stage, and tens of emitting stages within a laser core, this translates to hundreds of thin layers per device.

The material system used for making QCLs ranges from relatively mature GaInAs/AlInAs on InP substrate (InP based system) or AlGaAs/GaAs on GaAs substrate (GaAs based system) to relatively immature InAs/AlSb on InAs substrate (InAs based system). In addition, Silicon based system, e.g., Si/SiGe, and nitride based system, e.g., GaN/AlN are under theoretical investigation as host systems for intersubband lasing.

To form a comb-shaped conduction band profile as in Fig. 7.2, we need to alternate two materials with different conduction band energies. The conduction band offset ΔE_c refers to this energy difference. As also shown in Fig. 7.2, the photon energy, i.e., the energy difference between the upper laser subband and the lower laser subband, is smaller than ΔE_c. Although the wavelength can be changed somewhat by changing the layer thickness to shift the subband positions, the shortest achievable wavelength for a certain material system is bounded by ΔE_c. In addition, if the upper laser subband is made too close to the top of the barrier, electrons can be thermally excited into the continuum, i.e., states above the barrier, causing leakage.

Using Vegard's law, which assumes the average lattice constant varies linearly with the alloy content, the composition of a $\mathrm{Ga}_x\mathrm{In}_{1-x}\mathrm{As}/\mathrm{Al}_y\mathrm{In}_{1-y}\mathrm{As}$ heterostructure can be chosen such that the lattice constant matches that of the InP substrate. This case is called lattice matched (Fig. 7.6(a)). It is easiest to grow crystals with the same lattice constant as the substrate. However, in a mismatched system there are many interesting effects that can occur before dislocations start to form.

A mismatched crystal is initially pseudomorphic. In other words, it tries to take the form of the substrate by deforming itself at the interface between itself and the substrate (Fig. 7.6 (b, c)). Most growth methods are two-dimensional, which results in compressive (tensile) distortion in one plane and tensile (compressive) distortion in the third dimension to compensate. This distortion is called strain, and it affects the band structure due to displacement of the atoms in the unit cell. This strain exists until a certain critical thickness, at which point the strain is relieved by the formation of dislocations.

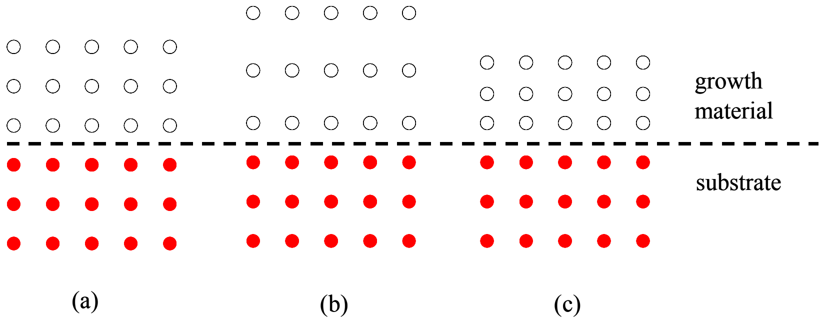


Fig. 7.6. Schematic diagram of lattice matched (a) and strained (b, c) heterostructures.

In order to limit the carrier escape from the top of the barrier for short wavelength designs and to extend the short wavelength limit, the conduction band offset can be increased by using a strain-balanced $Ga_xIn_{1-x}As/Al_yIn_{1-y}As$ heterostructure pseudomorphically grown on an InP substrate, where x and y are chosen such that the compressive strain in the wells is balanced by equal and opposite tensile strain in the barriers [Evans et al., 2004]. Due to the absence of net strain, the strain-balanced QCL wafer can be grown free of misfit dislocations up to a certain degree. The conduction band offset of the strain-balanced heterostructure can be increased by at least 50% more than the lattice matched $Ga_{0.47}In_{0.53}As/Al_{0.48}In_{0.52}As$ heterostructure, as shown in Fig. 7.7.

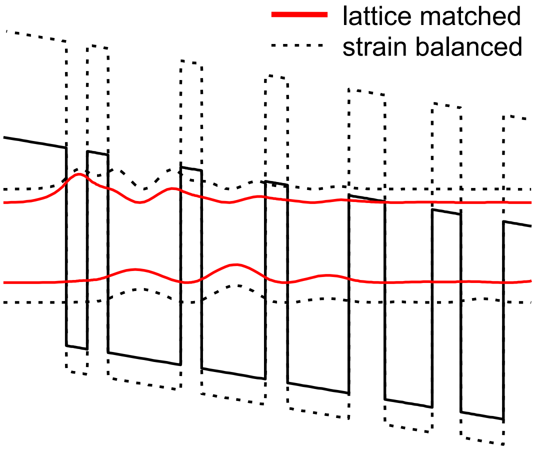


Fig. 7.7. Conduction band profiles and the corresponding upper and lower laser levels of lattice matched $Ga_{0.47}In_{0.53}As/Al_{0.48}In_{0.52}As$ and strain balanced $Ga_{0.33}In_{0.67}As/Al_{0.65}In_{0.35}As$.

To design a strain balanced QCL core heterostructure, the net strain of the system:

Eq. (7.23) $$\varepsilon_{\perp} = \sum_{i=1}^{N} h_i \varepsilon_{\perp i} \Big/ \sum_{i=1}^{N} h_i$$

needs to be minimized, where h_i and $\varepsilon_{\perp i}$ are the thickness and perpendicular strain of each layer, with $\varepsilon_{\perp i}$ defined by:

Eq. (7.24) $$\varepsilon_{\perp i} = \frac{a_{\perp i}}{a_i} - 1$$

where a_i is the equilibrium lattice constant, i.e., the lattice constant from Vegard's law and $a_{\perp i}$ is the lattice constant in the direction perpendicular to the substrate and is given by:

Eq. (7.25) $$a_{\perp i} = a_i \left[1 - D\left(\frac{a_{//}}{a_i} - 1 \right) \right]$$

with $a_{//}$ denoting the in-plane lattice constant and $D = \frac{\varepsilon_{\perp}}{\varepsilon_{\parallel}}$ relating the ratio of the perpendicular strain to the strain in the plane of the substrate. In the (001) direction, we can define D in terms of the diagonal (c_{12}) and off diagonal (c_{11})components of the elastic modulus for a cubic crystal:

Eq. (7.26) $$D = 2\frac{c_{12}}{c_{11}}$$

Since the pseudomorphic system implies that the in-plane lattice constant $a_{//}$ takes the lattice constant of the substrate a_{sub} (Fig. 7.6), we can calculate the perpendicular lattice constant of each material (Eq. (7.25)), and calculate the composition of materials required to minimize the net strain.

7.3.2. Laser waveguide

A functional QCL also needs a waveguide. The QCL waveguide has to provide both optical confinement and efficient heat extraction in continuious wave (CW) operation. In some cases these functions are in conflict with each other. For a higher confinement factor, one could put $Ga_{0.47}In_{0.53}As$ on both sides of the laser core due to its relatively higher refractive index than InP. However, being a ternary alloy, the thermal conductivity of $Ga_{0.47}In_{0.53}As$ is much lower than that of the binary InP. The same argument applies for putting more QCL stages inside the core. More stages gives better confinement but poorer thermal performance. Therefore, one has to optimize the laser waveguide according to the intrinsic properties of the laser core. Fig. 7.8 shows an example of the waveguide structure and the corresponding optical mode profile in the transverse direction (growth direction) of a QCL at a lasing wavelength of 4.5 μm. The absence of ternary material in the waveguide structure greatly enhances the heat removal in the transverse direction, which allows for better CW performance with only a small sacrifice of the confinement factor.

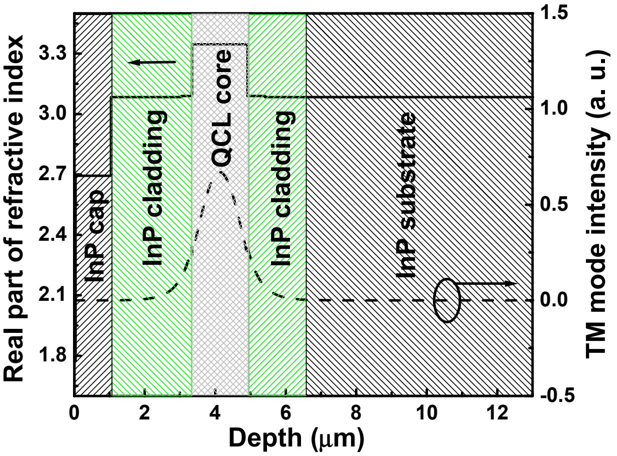


Fig. 7.8. Plot of the real component of the refractive index profile and the corresponding fundamental TM mode profile for a QCL waveguide structure.

Furthermore, the doping of the waveguide needs to be optimized. A lower doping level near the waveguide core favors a smaller optical loss (lower free carrier absorption) but also gives a poorer electrical conductivity, which leads to excess resistance. At the position of the cap layer, far from the laser core, where the overlap with the optical field is minimal, the doping can be made very high without introducing appreciable optical loss. This highly doped layer is necessary as it isolates the waveguide mode from coupling to the lossy surface plasmon mode at the metal-semiconductor interface. In the case of Fig. 7.8, starting from the InP substrate, the rest of the waveguide consists of a 1 μm low-doped (Si: ~2×10^{16} cm^{-3}) InP lower cladding layer, a 1.6 μm QCL core, a 3 μm low-doped (Si: ~2×10^{16} cm^{-3}) InP upper cladding layer, and a 1 μm high-doped (Si: ~1×10^{19} cm^{-3}) InP cap layer.

7.4. Making a quantum cascade laser

Even with all the design information available, the actual epitaxial growth has to be able to accurately recreate the intended structure. This is achieved by material characterizations. If the material quality is good, the QCL wafer needs to be processed and packaged to produce individual devices.

7.4.1. Epitaxial growth and material characterization

At present, both molecular beam epitaxy (MBE) and metalorganic chemical vapor deposition (MOCVD) have successfully grown high quality QCLs. Post growth characterizations need to be performed for each growth to ensure high material quality. (see Chapter 1 In particular, X-ray diffraction is the main tool to characterize the layer thickness and composition for a QCL structure.

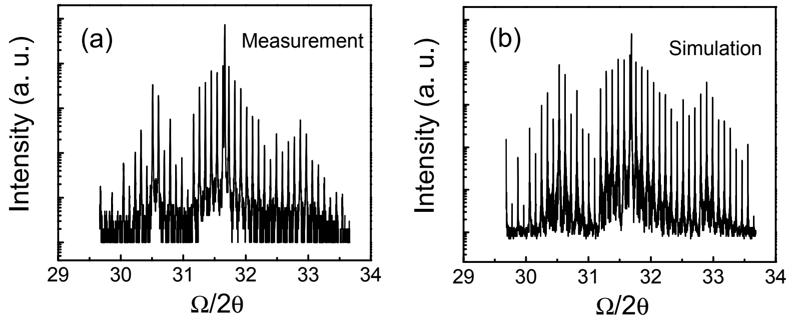


Fig. 7.9. Experimental measurement (a) and computer simulation (b) of the X-ray diffraction for a 30-stage strain-balanced QCL structure. The regularly spaced peaks are the signature of a periodic structure. The material quality is reflected by the width of the peaks. The fine position and the relative intensity of these peaks are related to the layer thicknesses and material compositions of the QCL heterostructure.

Shown in Fig. 7.9 is a comparison of the simulated and the measured X-ray diffraction curves. The close resemblance indicates precise recreation of the design parameters. In addition, optical characterization, such as photoluminescence, is used to probe the band structure. Varying layer composition or quantum well width can be detected by changes in the luminescence wavelength or lineshape. Electrical characterization is also very important. Hall effect or capacitance-voltage techniques can be used to identify doping levels and dopant distributions, which are critical to device operation.

7.4.2. Processing and packaging

Two kinds of processing methods are widely adopted for QCL fabrication, i.e., the double channel (DC) and buried ridge (BR) geometry. (Fig. 7.10)

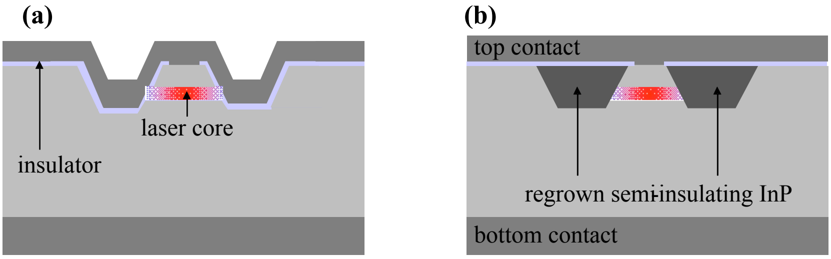


Fig. 7.10. Schematic diagram of double channel (a) and buried ridge (b) geometries.

In both geometries, the planar active region is isolated by channels etched into the wafer. This defines the lateral extent of optical and electrical propagation within the structure. For the double channel geometry, the whole surface is then covered by a dielectric insulator with the exception of a narrow window opened on top of the ridge allowing for selective electrical injection. For the buried ridge geometry, the channels are selectively filled with semi-insulating InP instead of a traditional dielectric such as SiO_2. This allows for lower thermal and refractive index mismatch at the core/channel interface.

After standard metallization and lapping/polishing, the QCL wafer is cleaved into laser bars with the desired cavity length. For better heat removal and current injection, a laser bar needs to be packaged, i.e., bonded to a heatsink. There are two bonding methods, namely the epilayer-up bonding and the epilayer-down bonding, as shown in Fig. 7.11. The epilayer-up method is the simplest, in which the substrate side of the laser bar is directly bonded (soldered) to the heatsink. In the epilayer-down bonding method, however, the epilayer side is first attached to a submount, which is made of materials with high thermal conductivity, such as AlN or diamond, and the submount is then bonded to the heatsink. The latter technique gives better thermal conductance for the package as the laser core is closer to the heatsink. However, the technique is more difficult as there is no tolerance for solder squeeze-out or solder voids under the laser core. As such, specialized submount preparation and bonding technology is required.

Finally, an illustration for the structure of a QCL at different magnifications is depicted in Fig. 7.12.

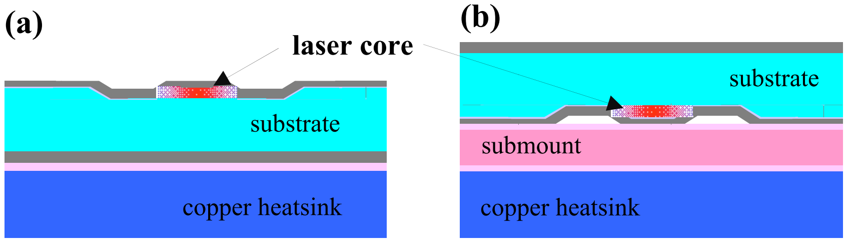

(a)
laser core
substrate
copper heatsink
(b)
substrate
submount
copper heatsink

Fig. 7.11. Schematic diagram of epilayer-up (a) and epilayer-down (b) bonding.

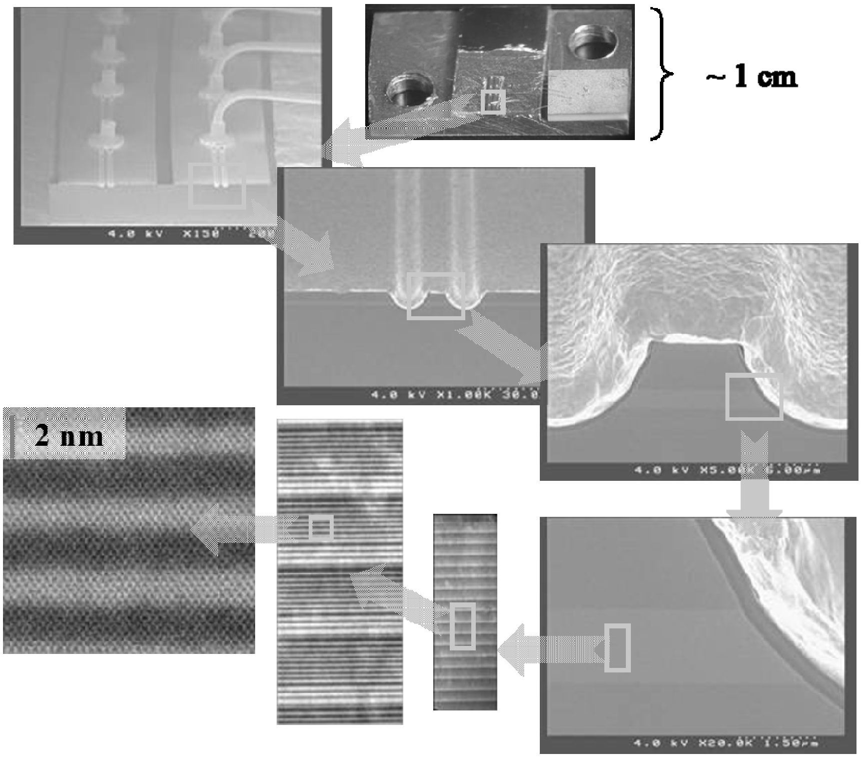

~ 1 cm
2 nm
4.0 kV X150
4.0 kV X1.00K 30.0
4.0 kV X5.00K 6.00µm
4.0 kV X20.0K 1.50µm

Fig. 7.12. Illustration of various size scales relevant to the quantum cascade laser. At the top is a packaged device. In the middle is the waveguide cross-section as imaged by a scanning electron microscope. On the lower left are some of the individual layers of the injector region as imaged by a transmission electron microscope. [Reproduced with permission from Optics and Photonics News July 2008, M. Razeghi, "The Quantum Cascade Laser, a Versatile and Powerful Tool," pg. 46, Copyright 2008, Optics Society of America (OSA).)]

7.5. Device performance

Basic device performance of a QCL include the power-current-voltage characteristics (PIV), temperature dependent characteristics, and wall plug efficiency (WPE). For most applications, higher output power and higher wall plug efficiency are always preferred. Even for those applications that do not require high optical power, a QCL with high wall plug efficiency can keep the input electrical power at a low level, hence alleviate the burden of driving and cooling the device during operation. In addition, the emission spectrum and far field are also very important. A QCL with a narrow emitting spectrum and a diffraction limited far field, i.e., a single mode QCL, is preferred for most applications. A single mode spectrum means a well-defined wavelength. When used as a probing source, a QCL with a narrow emission spectrum minimizes cross talk, hence allows for high selectivity. A diffraction limited far field gives the maximum brightness for the case with no external lenses. It also allows for focusing the laser beam with standard lenses.

7.5.1. Power-current-voltage characteristics

The output power (P) and operating voltage (V) of a QCL as a function of the driving current (I) are the basic device performance characterizations. Due to the fact that the internal heating plays an important role in the QCL operation, the testing condition and the device packaging need to be specified for the PIV curves. For most cases, QCLs are tested with a thermoelectric cooler (TEC), which actively regulates the heatsink temperature. This temperature is one of the most important aspects for the testing condition. In addition, a QCL can be driven either in pulsed mode or in continuous wave (CW) mode. In pulsed mode operation, the laser is repetitively turned on and off by an external signal. An important parameter called the duty cycle represents the percentage of time of which the laser is on during one repetition. For example, if the repetition rate of the signal is 100 kHz and the pulse width is 500 ns, this gives a duty cycle of 5%. When the duty cycle reaches 100%, i.e., the laser is on all the time, it operates in CW mode.

Aspects for the device packaging include the processing method, e.g., double channel or buried ridge, the laser core geometry, i.e., the ridge width and the cavity length, and the bonding method, i.e., epilayer-up or epilayer-down. It is necessary to specify the submount material, e.g., AlN or diamond, if epilayer-down is used. In addition, it is also necessary to specify the facet condition, i.e., uncoated or coated. An uncoated facet is formed by the natural cleavage plane with a reflectivity of about 26% in the mid-infrared region for the GaInAs/AlInAs material system. The facet reflectivity can be modified by coating with metal or dielectric. The reflectivity of metal, e.g., Au, in the mid-infrared region is nearly unity, which makes it an excellent candidate for high reflectivity (HR) coatings. Although more complicated than coating with metal, dielectrics can also form HR coatings. By stacking dielectric materials of different refractive indexes with proper thicknesses, anti-reflectivity (AR) coatings can also be obtained [Nguyen et al. 2006]. In general, an uncoated device refers to a device with both facets left uncoated and an HR coated device refers to a device with one facet coated with metal and the other one left uncoated. For an uncoated device, the same optical power is assumed to come out of both facets and the power is only measured from the front facet. For a HR coated device, the front facet is always the uncoated facet and the power coming out of the back facet (HR coated) is assumed to be negligible.

As an example, Fig. 7.13 shows the PIV curves for a 4.7 μm QCL. The device is tested at 298 K in pulsed mode with a duty cycle of 5% and a pulsed width of 500 ns. The ridge width (W) and cavity length (L) are 10.6 μm and 3 mm, respectively. The wafer is processed with the double channel method and the laser bar is epilayer-up bonded. The back facet is HR coated.

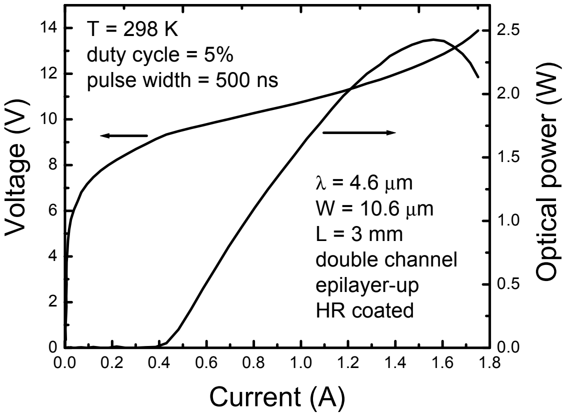


Fig. 7.13. Power-current-voltage characteristics of a 4.6 μm QCL.

From the PIV characteristics, several important device operation parameters can be extracted, including the threshold current I_{th}, threshold voltage V_{th}, maximum power P_{max}, rollover current I_{max}, slope efficiency η_s, and differential resistance R. The threshold current is the current at which the output power starts to increase substantially. The threshold voltage is the operating voltage at the threshold current. The peak of the PI curve gives the maximum power and the rollover current. The slope of a linear fit to the PI (VI) curve above the threshold current and below the rollover current gives the slope efficiency (differential resistance). For the case of Fig. 7.13, the following device operation parameters can be obtained, $I_{th} = 0.43\text{ A}$, $V_{th} = 9.3\text{ V}$, $P_{max} = 2.43\text{ W}$, $I_{max} = 1.56\text{ A}$, $\eta_s = 2.94\text{ W/A}$, $R = 2.44\ \Omega$.

Since the area of the current cross section $A = WL$, we can normalize the threshold current and the differential resistance, resulting in the threshold current density $J_{th} = I_{th} / A$ and the area resistance RA. For the case of Fig. 7.13, we have $J_{th} = 1.35\text{ kA/cm}^2$, $RA = 0.78\text{ m}\Omega \cdot \text{cm}^2$.

7.5.2. Temperature dependent characteristics

One of the advantages of QCLs over the conventional interband lasers is that their performances are relatively insensitive to changes in temperature. However, if we change the temperature over a wide range, we can still observe trends in overall performances.

In general, the change of the PI curve is more pronounced than the change of VI curve when the temperature changes, as shown in Fig. 7.14.

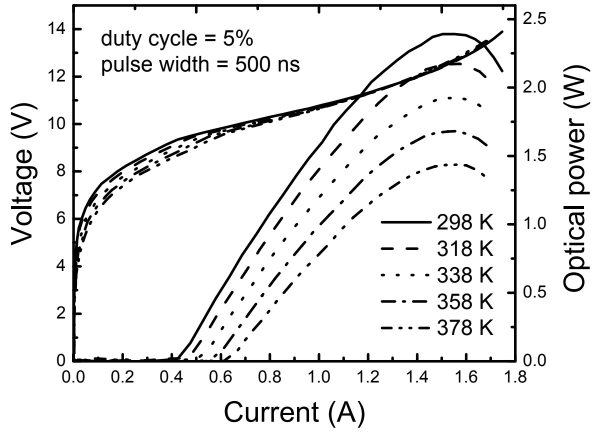


Fig. 7.14. Temperature dependent power-current-voltage characteristics.

The threshold current density increases and the slope efficiency decreases as a function of the heatsink temperature, as shown in Fig. 7.15. In particular, the threshold current density as a function of temperature is usually fitted by an exponential expression $J_{th} = J_0 \exp(T / T_0)$, with the characteristic temperature T_0 representing the insensitivity of the threshold current density to the temperature. A higher T_0 means a better device.

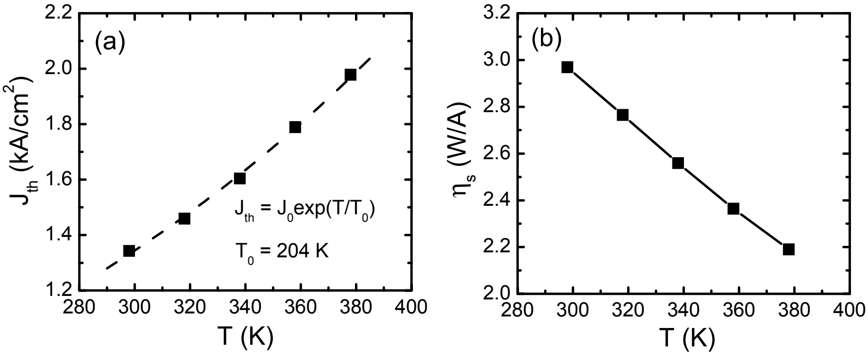


Fig. 7.15. Temperature dependent threshold current density (a) and slope efficiency (b).

The main difference between the pulsed mode and continuious wave (CW) mode operation is the temperature of the laser core. In pulsed mode, at sufficiently low duty cycles, the temperature of the laser core is about the same as that of the heatsink. However, in CW mode, the temperature of the laser core is expected to be much higher than that of the heatsink due to the self-heating effect. As shown in Fig. 7.16, if we overlay the PIV curves in CW mode with those in the pulsed mode, we can see that the CW curves intercept with multiple pulsed curves at different temperatures. From these interceptions, the core temperature at different driving currents can be estimated for the CW operation. For example, the core temperature is about 350 K at threshold and higher than 378 K at the maximum current.

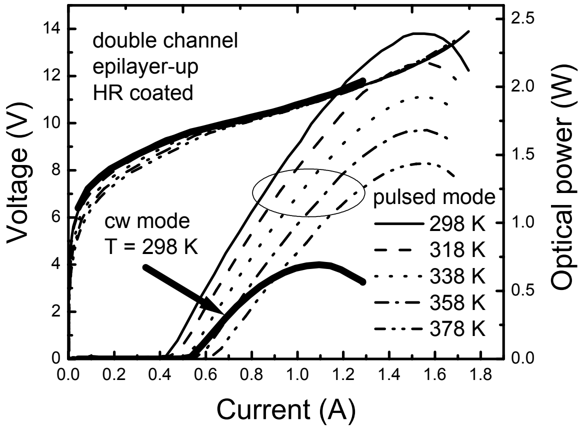


Fig. 7.16. Power-current-voltage characteristics of a 4.6 μm QCL operating in CW mode at a heatsink temperature of 298 K (thick lines) to the pulsed mode characteristics at different temperatures (thin lines).

7.5.3. Wall plug efficiency

The wall plug efficiency (WPE) is defined as the portion of the injected electrical power converted into the optical power:

Eq. (7.27) $$\eta_w = \frac{P}{IV}$$

where P and V can be replaced by:

Eq. (7.28) $$P = \eta_s (I - I_{th})$$

Eq. (7.29) $$V = V_{th} + (I - I_{th})R$$

The slope efficiency η_s in Eq. (7.28) can be replaced by Eq. (7.22). After some algebra, we can factorize the WPE into four component sub-efficiencies:

Eq. (7.30) $$\eta_w = \eta_i \eta_o \eta_v \eta_e$$

with the following expression for the optical efficiency η_o, voltage efficiency η_v, and electrical efficiency η_e:

Eq. (7.31) $$\eta_o = \frac{\alpha_m}{\alpha_m + \alpha_w}$$

Eq. (7.32) $$\eta_v = \frac{N\hbar\omega}{e}\frac{1}{V_{th}}$$

Eq. (7.33) $$\eta_e = \frac{I - I_{th}}{I\left[1 + \frac{R(I - I_{th})}{V_{th}}\right]}$$

The factorization is done in such a manner that each composing sub-efficiency is unitless and physically meaningful. In addition, the current-dependent nature of the WPE is reflected in the electrical efficiency only. Note that obtaining Eq. (7.30) – Eq. (7.33) only requires two assumptions, namely, a constant η_i and a constant R . The electrical efficiency can also be rewritten in terms of the current density and threshold current density:

Eq. (7.34) $$\eta_e = \frac{J - J_{th}}{J\left[1 + \frac{RA(J - J_{th})}{V_{th}}\right]}$$

In order to get an idea of all the sub-efficiencies, several device performance parameters are needed, including J_{th}, V_{th}, RA , α_w, and η_i . The first three of them are readily available from just one current-voltage-power measurement to the device, e.g., Fig. 7.13. According to Eq. (7.22), we can separate η_i and α_w by changing α_m, which leads to the extraction of the internal quantum efficiency and the waveguide loss simultaneously, according the following procedure.

Eq. (7.22) can be rearranged into:

Eq. (7.35) $$\frac{1}{\eta_{ext}} = \frac{\alpha_w}{\eta_i}\left(\frac{1}{\alpha_m}\right) + \frac{1}{\eta_i}$$

with the external quantum efficiency η_{ext} defined by:

Eq. (7.36) $$\eta_{ext} = \eta_s \frac{e}{N\hbar\omega}$$

If we perform slope efficiency measurements for a number of devices with different mirror losses, either by changing the cavity length or optical coating to the facet, via:

Eq. (7.37) $$\alpha_m = \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2} \right)$$

where R_1 and R_2 are facet reflectivities, we can prepare the data according to Eq. (7.35) and plot the inverse of the external quantum efficiency versus the inverse of the mirror loss, and then perform a linear fit. As a result, the internal quantum efficiency follows as the inverse of the y-intercept of the fit. The waveguide loss can be obtained from the slope of the fit with the known internal quantum efficiency. Fig. 7.17 shows an example of this procedure, where the fitting is performed upon pulsed data from 9 QCLs, covering four different mirror losses with 3 mm uncoated, 3 mm HR-coated, 4 mm uncoated, and 4 mm HR-coated devices. In some cases, multiple devices with the same mirror loss are included to show the uncertainty of measuring the external quantum efficiency, e.g., 5 devices are tested for the 3 mm HR-coated case. (Some of them overlap.)

Due to the inter-device-comparison nature of the method, data consistency is the most important factor in limiting the error. We suggest power measurement using a calibrated thermopile in close proximity to the laser facet, which avoids power calibration due to low collection efficiency when using photodetectors and lenses. As mentioned, the parameters are best extracted for pulsed measurements, where internal heating is negligible, but there is sufficient average power to be measured with a thermopile detector. This balance means more reliable power measurement for lasers with higher slope efficiency.

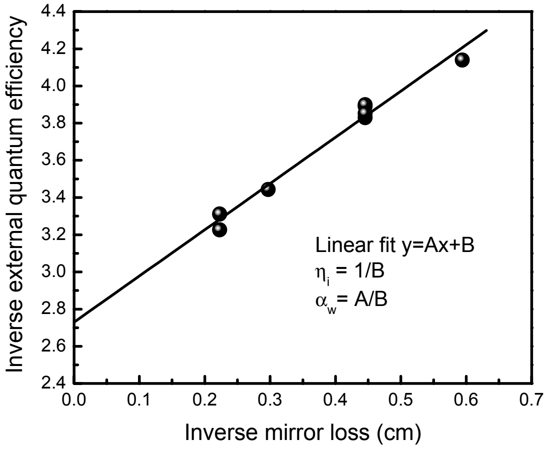


Fig. 7.17. Illustration for the method of experimentally extracting the internal quantum efficiency and the waveguide loss.

7.5.4. Spectra and far field

For a QCL with a Fabry Perot (FP) cavity, the principle determining the characteristics of the spectrum and far field follows the discussion in Chapter 6 since the QCL falls in the category of the semiconductor laser.

As discussed earlier, from the design point of view, the wavelength of a QCL is determined by the energy difference between the upper and lower laser levels. In a real device, however, both the upper and lower laser levels are broadened due to various scattering mechanisms. As a result, the material gain spans a range of wavelengths in the vicinity of the design wavelength. What eventually appear in the lasing spectrum are those that selected by the feedback property of the cavity. For an FP cavity, we tend to have multi-mode spectrum. Single mode spectrum can be achieved with the incorporation of distributed feedback (DFB) mechanism. (Fig. 7.18 (a))

The far field of a semiconductor laser is the Frauenhofer transform of the near field, so a single mode near field gives a single mode far field and a multi-mode near field gives a multi-mold far field. There exist many near field modes for a QCL waveguide. The optical losses are different among them. If there is only one mode (usually the fundamental mode) that has an optical loss significantly smaller than others, the laser action will occur in that mode and the far field will be single mode. In other cases, when there are many modes that have similar low optical losses, the far field will be multi-mode. For a typical mid-infrared QCL, the thickness of the laser core is comparable to the wavelength inside the material, hence the margin of the optical loss between the fundamental mode and the second order mode is very big. As such, the far field in the transverse direction is always single mode. In the lateral direction, however, the ridge width can vary from several micrometers to hundreds of micrometers and the far field is not always single mode. For QCLs with relatively narrower ridge widths, the far field is usually single mode. As the ridge width increase, more and more lateral modes show similar optical losses and the far field becomes multi-mode. (Fig. 7.18 (b))

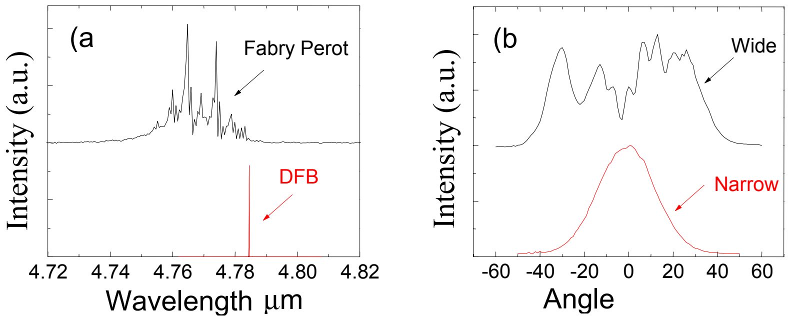


Fig. 7.18. Spectra (a) and far field (b) characteristics of a QCL. The ridge width of "wide" and "narrow" devices in (b) are 50 µm and 10 µm, respectively.

7.6. Wall plug efficiency optimization

The total wall plug efficiency (WPE) can be improved for manipulations to the device that enhance some of the four sub-efficiencies without sacrificing the others. Otherwise a compromise needs to be made. The internal quantum efficiency can only be improved by investigating the design of the QCL core heterostructure, however that is beyond the scope of this text. For the same QCL core heterostructure, there are strategies that can be pursued to increase other sub-efficiencies, with the same internal quantum efficiency.

7.6.1. Electrical contact resistance

One of the easiest manipulations that can be done to the device is to change the contact metallurgy. Since the metal contact is in series with the laser, it does not influence the threshold current. Because it is far away from the laser core, it has little effect on the waveguide loss and no effect on the internal quantum efficiency. Its influence is limited to the threshold voltage and differential resistance only. Hence manipulating the contact only changes the voltage efficiency and electrical efficiency. Preferably, both the top and bottom contact should be ohmic with negligible contact resistances.

Most of the InP based QCL wafers were grown on n^- InP substrate (Sn: 2–3 × 10^{17} cm^3) with a highly doped n^+ InP (Si: 1–2 × 10^{19} cm^3) cap layer to finish the structure. The Ti/Au top and bottom contact had been adopted by many groups. From the contact resistance study we found that the bottom Ti/Au was problematic, whereas the top Ti/Au contact was acceptable. To confirm and solve the bottom contact problem, we reprocessed one of the QCL wafers that had been processed with the Ti/Au bottom contact with a new Ge/Au/Ni/Au bottom contact. Shown in Fig. 7.19 is the normalized comparison. By changing the bottom contact from Ti/Au to Ge/Au/Ni/Au, we are able to reduce the threshold voltage and the area resistance by 1.7 V and 0.5 mΩ·cm^2, respectively. An explanation for the low contact resistance for the Ge containing contact on low-doped InP surface is that Ge diffuses into InP as a donor atom, resulting a thin layer of high-doped InP at the metal-semiconductor interface, which is then depleted to form a low resistance tunnel contact. As a result, the voltage efficiency was increased from 59% to 67% and the electrical efficiency was increased from 41% to 42% at a current density of 2.5 kA/cm^2. Although the enhancement to the electrical efficiency is small, the voltage efficiency is increased by about 13%, which means about 13% enhancement to the WPE can be achieved by changing the bottom contact.

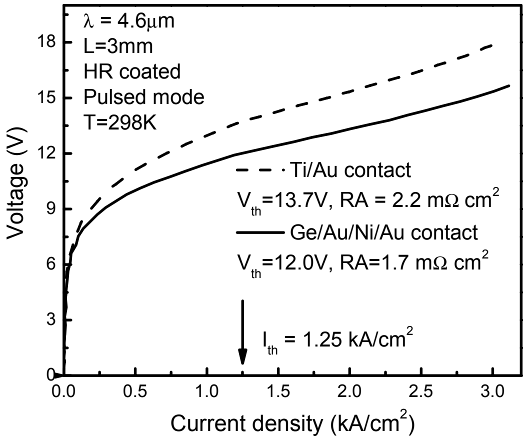


Fig. 7.19. Current-voltage characteristics comparison of two 4.6 μm QCLs processed from the same wafer with different bottom contacts. [Reproduced with permission from IEEE Journal of Selected Topics in Quantum Electronics Vol. 15, M. Razeghi, "High performance InP-Based Mid-IR Quantum Cascade Lasers," fig. 8, pg. 947, Copyright 2009 IEEE.]

7.6.2. Waveguide geometry

The role of the waveguide, related to wall plug efficiency (WPE), is mainly through the waveguide loss. Although changing the waveguide loss also influences the threshold current density and the threshold voltage, the relative influence to the electrical efficiency and voltage efficiency is much smaller than the influence to the optical efficiency for typical QCLs.

Strictly speaking, the QCL core is part of the waveguide, and the geometry of the core plays an important role in determining the properties of the waveguide. For the same QCL core geometry, the surrounding material also influences the properties of the waveguide. For the double channel processed QCLs, the waveguide in the lateral direction is bounded by two etched channels, which are covered by an insulator and metals. Although the combination of insulator and metals provides excellent lateral confinement, they also introduce extra waveguide loss, which depends on their overlap with the optical mode. Therefore, we expect to see a change of waveguide loss if the overlap is changed. An easy way to change the overlap is to change the separation of the two channels, i.e., to change the ridge width. Simulation suggests a wider ridge width will show a lower waveguide loss. Experimentally, as shown in Fig. 7.20, waveguide losses of 1.0 ±0.1 cm^{-1} and 0.7 ±0.1 cm^{-1} are obtained from the mirror loss dependent slope efficiency measurement for the double channel processed QCLs with narrow and wide ridge widths, respectively. For the 3 mm HR-coated device, the reduction of a waveguide loss from 1.0 cm^{-1} to 0.7 cm^{-1} would translate to an increase of optical efficiency from 69% to 76%, which accounts for a 10% enhancement to the WPE

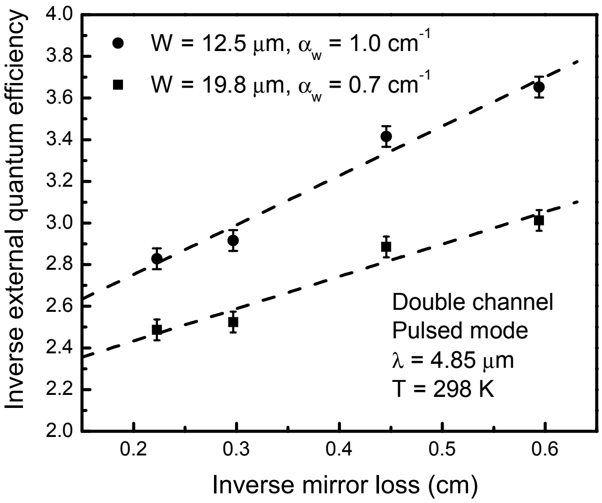


Fig. 7.20. Waveguide loss measurement for 4.85 μm QCLs with double channel geometry in narrow and wide ridge widths. [Reproduced with permission from IEEE Journal of Selected Topics in Quantum Electronics Vol. 15, M. Razeghi, "High performance InP-Based Mid-IR Quantum Cascade Lasers," fig. 10, pg. 947, Copyright 2009 IEEE.]

Although a wider ridge width with a double channel geometry favors a higher pulsed WPE, it is not suited for high temperature continuious wave (CW) operation due to inefficient lateral heat removal. As evidence, the two HR coated narrow-ridge-width devices in Fig. 7.20 exhibited CW operation at room temperature, but none of the wide-ridge-width devices were able to lase in CW mode at room temperature, even though they had a higher pulsed WPE. For this particular QCL wafer, experimental results suggest that an even narrower ridge width is the direction for better CW performance. But this will come at the expense of a smaller pulsed WPE, which is the upper limit of the CW WPE for the same device operating at the same heatsink temperature. As the DC ridge width continues to decrease, the reduction of this upper limit will eventually overwhelm the benefit of heat removal.

The buried ridge geometry, however, allows for small waveguide losses at very narrow ridge widths, hence it is advantageous both optically and thermally. In this case, filling the channels with low loss semi-insulating InP removes the presence of insulator and metal on the sides of the ridge, therefore decreasing the waveguide loss. In addition, the regrown InP also offers a good lateral heat dissipation path, due to its superior thermal conductivity compared to the insulator in the double channel geometry.

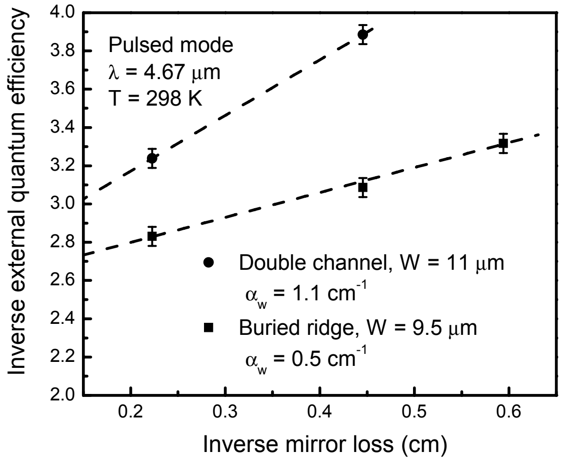


Fig. 7.21. Waveguide loss measurement for 4.67 μm QCLs with double channel (11 μm ridge width) and buried ridge (9.5 μm ridge width) geometries. [Reproduced with permission from IEEE Journal of Selected Topics in Quantum Electronics Vol. 15, M. Razeghi, "High performance InP-Based Mid-IR Quantum Cascade Lasers," fig. 11, pg. 948, Copyright 2009 IEEE.]

Shown in Fig. 7.21 is the experimental measurement for the waveguide loss for the same QCL wafer processed in both double channel (DC) and curried ridge (BR) geometry. Although the ridge width of the BR geometry is narrower, it only exhibits a waveguide loss of 0.5 ±0.1 cm^{-1}, compared with 1.1 ±0.1 cm^{-1} for the DC geometry with a wider ridge width. For the 3 mm HR-coated device, this will increase the optical efficiency from 67% to 82%, which translates to 22% enhancement to the WPE.

The combination of using the Ge/Au/Ni/Au bottom contact and buried ridge geometry will provide about a 38% total enhancement to the pulsed WPE.

7.6.3. Bonding method

The threshold power density for mid-infrared QCLs is typically on the order of 10 $\mathrm{kW/cm}^2$, which is much higher than that of an interband laser. Therefore, the thermal packaging for high temperature CW operation is much more demanding for QCLs. As a result, thermal considerations are incorporated throughout the material growth, wafer processing, and laser bonding. In general, the thermal management for the QCL can be categorized into concerns about the heat source, i.e., the laser core, and the thermal conductivity of the surrounding material. The first concern favors a device that has a smaller threshold power density and higher slope efficiency in pulsed mode operation. In addition, the thermal conductivity and the geometry of the core itself are also very important. The second concern favors a device that uses materials with high thermal conductivity along the path of the heat flow.

Once a QCL wafer is processed into a certain waveguide geometry, the surrounding material in the vicinity of the laser core cannot be changed. However, due to the structural asymmetry in the transverse direction, one can change the heat path by mounting the laser bar in different orientations, i.e., epilayer-up or epilayer-down, as discussed earlier. For epilayer-up bonding, the upward heat flux out of the laser core is often enhanced by the use of a thick (~5 µm) electroplated gold layer on top of the device surface, which spreads the heat laterally and eventually transfers most of the heat back downward to the substrate at passive regions, since the convection at the gold-air interface is very inefficient due to a small surface area. The downward heat flux directly out of the laser core, and that from the electroplated gold, has to travel through a fairly thick (~100 µm) InP substrate before reaching the copper heatsink.

Because the thermal conductivity of InP (~70 $\mathrm{W{\cdot}m^{-1}{\cdot}K^{-1}}$) is much smaller than that of the copper (~400 $\mathrm{W{\cdot}m^{-1}{\cdot}K^{-1}}$), the thick substrate in the epilayer-up bonding method is essentially a thermal barrier between the laser core and heatsink. In the epilayer-down bonding method, however, this thermal barrier is replaced by a submount, which can be made of materials with higher thermal conductivity than that of InP, e.g., AlN

(~170 W·m^{-1}·K^{-1}) or diamond (~1800 W·m^{-1}·K^{-1}). For the case of the diamond submount, since its thermal conductivity is much higher than that of the copper, it also serves as a very good heat spreader, which spreads the heat coming out of a very narrow stripe (on the order of the ridge width) into a much bigger area before transferring this heat to the heatsink.

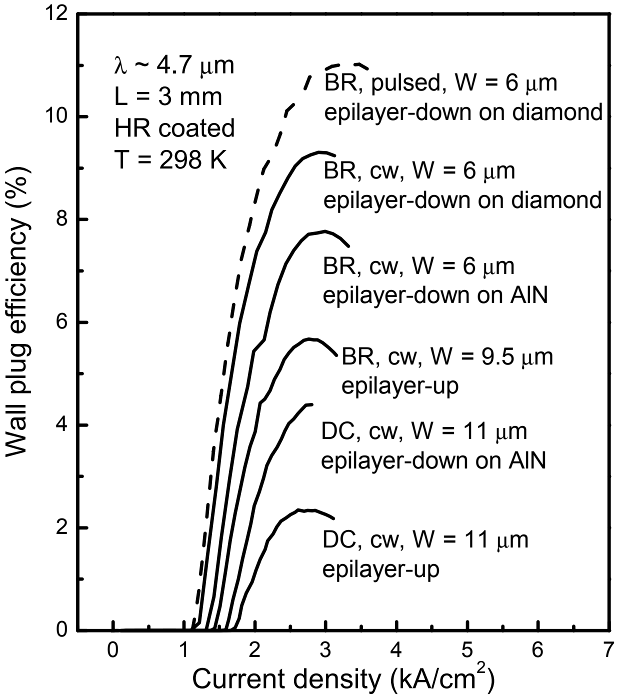


Fig. 7.22. Wall plug efficiency-current density characteristics for 4.7 μm QCLs with different thermal management tested at room temperature. [Reproduced with permission from IEEE Journal of Selected Topics in Quantum Electronics Vol. 15, M. Razeghi, "High performance InP-Based Mid-IR Quantum Cascade Lasers," fig. 12, pg. 949, Copyright 2009 IEEE.]

The experimental comparison for different bonding methods, as well as the influence of the waveguide geometry, is shown in Fig. 7.22, where all the devices are processed from the same QCL wafer. As expected, the combination of buried ridge geometry, narrow ridge width, epilayer-down bonding, and a diamond submount offers the closest continuious wave (CW) performance to the pulsed data. In general, the quality of the thermal management can be quantified by defining a thermal efficiency, which is the ratio of the maximum CW wall plug efficiency (WPE) to the pulsed WPE for the same device. The best thermal packaging in Fig. 7.22 offers 9.3% CW WPE out of 11% pulsed WPE, which translates to 84% thermal efficiency. Further enhancement to the thermal efficiency is possible for an even narrower buried ridge geometry. Eventually there is an optimum ridge width for the highest CW WPE because the waveguide loss increases due to the reduction of the optical confinement factor as the ridge width gets narrower. For the case in Fig. 7.22, this optimum ridge width is not greater than 6 μm.

7.7. Power scaling

The optimization of the total output power is different from the optimization of wall plug efficiency (WPE). As an efficiency measure, WPE is non-scalable and bounded between 0 and 1, whereas the output power is scalable with the size of the emitting core.

For pulsed mode operation, the peak output power is nearly proportional to the ridge width, whereas the dependence of WPE to the ridge width is very weak, as shown in the inset of Fig. 7.23, where the pulsed mode WPE and maximum output power for a number of devices with different ridge widths processed from the same wafer in the double channel geometry is presented. For devices with a ridge width smaller than 50 μm, the WPE increases with the ridge width, mainly due to the decrease of waveguide loss. Above 50 μm we start to see a decreasing of the WPE. The internal heating in pulsed mode and high order transverse mode oscillation for extremely wide ridge widths are possible reasons for this behavior. In fact, the far field for the 100 μm and 200 μm ridge-width devices are distinctively double lobed, which is a signature of a high order transverse mode. Despite the slight reduction in WPE, the maximum total peak power for the 200 μm ridge-width device still reaches 34 W at room temperature, which can be very attractive for applications that need a high power pulsed source.

Other ways to scale the output power include increasing the core doping, the number of QCL stages inside the core, and the cavity length. In general, power scaling can be categorized into changes made to the carrier density and the size of the core, which is 3-dimensional. It is understandable that increasing the core doping, number of QCL stages, and the cavity length will have a similar effect on the peak power in the pulsed mode as increasing the ridge width.

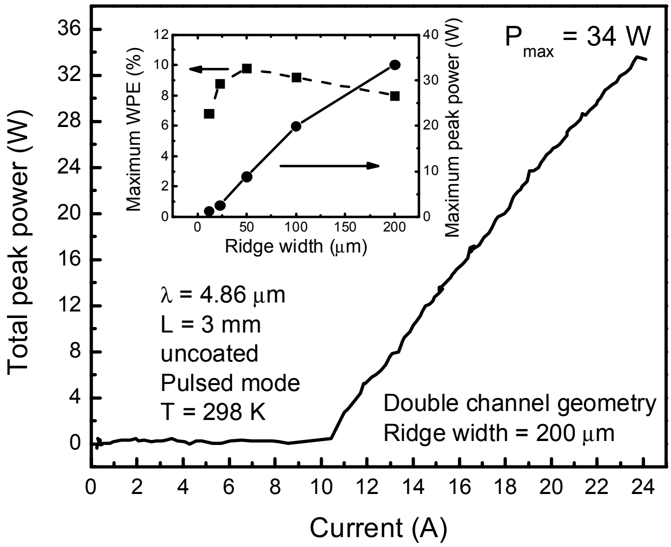


Fig. 7.23. Room temperature pulsed mode operation of a 4.86 μm QCL with a 200 μm ridge width. The inset shows the maximum WPE and peak power for a number of devices with different ridge widths. [Reproduced with permission from IEEE Journal of Selected Topics in Quantum Electronics Vol. 15, M. Razeghi, "High performance InP-Based Mid-IR Quantum Cascade Lasers," fig. 13, pg. 949, Copyright 2009 IEEE.]

Power scaling in continuious wave (CW) mode, however, is not as straightforward as in pulsed mode, because severe internal heating precludes the use of a very wide ridge width for QCLs with a typical pulsed WPE of about 10%. In addition, increasing the ridge width and the number of QCL stages is against the design rules for efficient heat removal. Increasing the core doping will result in a higher operating current density, but a higher power density leads to more heating. The only approach that has little effect on the thermal conductance or power density within the device is to increase the cavity length, but this is limited by practical difficulties of bonding a semiconductor laser with a very long cavity. As a result, power scaling in CW mode is sensitive to the balance of internal heating and the corresponding peak power in pulsed mode. With the development of more efficient QCL cores and advanced thermal management, this balance is expected to shift toward bigger emitter sizes.

The above discussion applies to the power scaling for a single emitter. If multiple QCLs are made into an array or even a stack of arrays, with an active cooling method, such as microchannel cooling, tens or even hundreds of watts of CW power is eventually possible. In this scenario, the maximum obtainable CW output power is only limited by the cooling capacity. For a fixed cooling capacity, the maximum obtainable CW power will increase dramatically for QCLs with higher WPE, as less heat needs to be dissipated for the same amount of output power.

7.8. Photonic crystal distributed feedback quantum cascade lasers

Photonic crystals are periodic optical structures that are designed to affect the motion of photons in a similar way that the periodicity of a crystal affects the motion of the electrons, namely by the introduction of periodic dispersions and forbidden energy gaps in the optical dispersion itself.

The far field and spectral properties of the broad area (ridge width greater than 50 μm) QCLs can be improved by the photonic crystal distributed feedback (PCDFB) mechanism, as illustrated in Fig. 7.24. In general, for Fabry Perot (FP) lasers, as the emitter width increases, multi-mode operation in both far field and spectrum happens due to poorer lateral and longitudinal mode distinguishability. The conventional distributed feedback (1D-DFB) mechanism is able to offer single mode operation in terms of the emitting wavelength, however, the far field is still multi-mode, because the feedback mechanism exists only in the longitudinal direction. Similarly, the angled grating DFB (α-DFB) allows for single mode operation in terms of the far field by introducing feedback in the lateral direction, but the spectrum is usually multi-mode. An approach that combines both the spatial beam quality of the α-DFB and the spectral beam quality of the 1-D DFB is the photonic crystal distributed feedback (PCDFB) laser, which generalizes the 1-D grating geometry to a 2-D pattern, allowing single mode operation both in far field and spectrum.

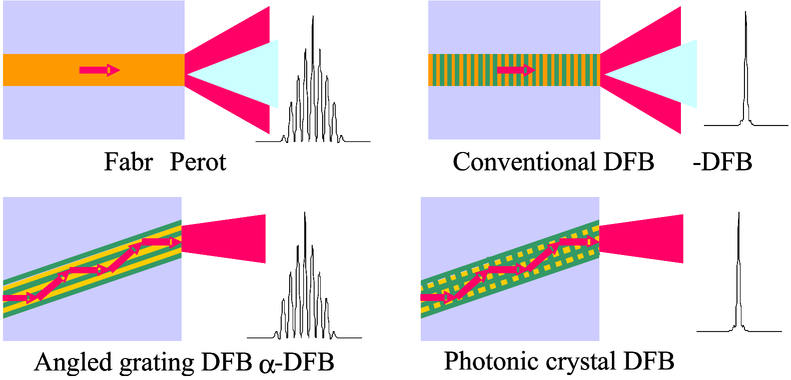


Fig. 7.24. Illustration of broad area QCLs with different feedback mechanisms. The far field and spectral characteristics are shown with shaded beams and solid curves, respectively.

The incorporation of 1D-DFB in QCLs has been demonstrated in a number of wavelengths with CW operation at room temperature [e.g., Yu et al. 2005, Darvish et al. 2006a, Darvish et al. 2006b]. Although α-DFB has not been demonstrated for QCLs, it is realized in mid-infrared Type-II lasers [Bartolo et al. 2000], which implies no obstacles for QCLs. Inspired by the theoretical work of applying PCDFB on QCLs [Vurgaftman et al. 2006], the electrically pumped PCDFB QCL has become an experimental reality [Bai et al. 2007].

7.8.1. Pattern design

The PCDFB QCL structure is shown in Fig. 7.25. It is basically a regular QCL with an additional patterned grating layer (GaInAs) just above the laser core. This grating layer provides both longitudinal and lateral feedbacks, which is responsible for single mode operation both spectrally and spatially.

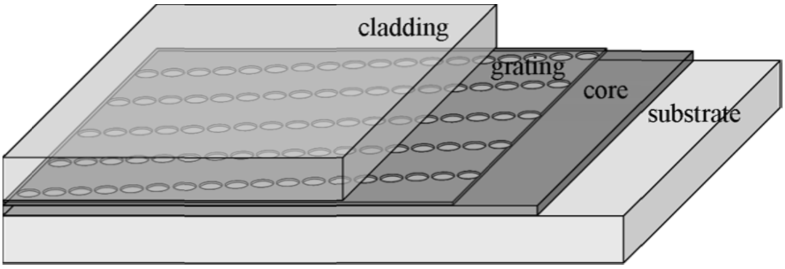


Fig. 7.25. Sketch of the PCDFB structure. [Reproduced with permission from Applied Physics Letters Vol. 91, Y. Bai, S.R. Darvish, S. Slivken, J. Nguyen, A. Evans, W. Zhang, and M. Razeghi, "Electrically pumped photonic crystal distributed feedback quantum cascade lasers," fig.1, pg. 141123-1, Copyright 2007 American Institute of Physics (AIP).]

One distinctive feature of the PCDFB (and α-DFB) pattern is that the patterned region is tiled with respect to the laser facet. This is necessary in order to align one of the primary propagation vectors with the facet normal. The PCDFB pattern must satisfy the 2D Brag condition:

Eq. (7.38) $$\Lambda_1 = \frac{\lambda}{2n_{eff}\sin\phi} m$$

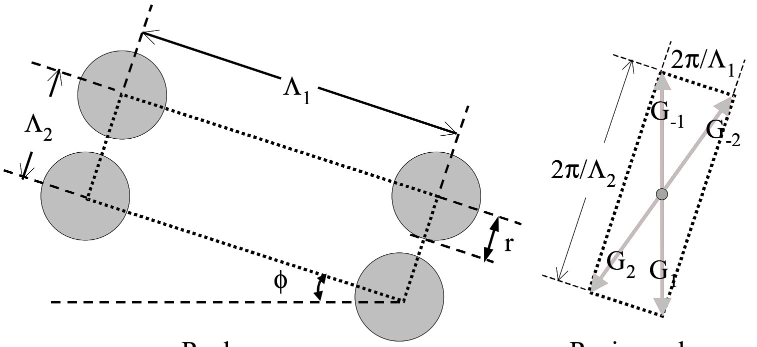
and

Eq. (7.39) $$\Lambda_2 = \frac{\lambda}{2n_{eff}\cos\phi} n$$

where Λ_1 and Λ_2 are the pattern periods in two orthogonal directions. n and m are integral numbers for the diffraction order. λ, n_{eff}, and ϕ are the wavelength, effective index and tilt angle, respectively. The first order coupling is the most efficient, hence we will set n and m to be 1. As such, the following relation holds between the tilt angle and the pattern periods:

Eq. (7.40) $$\tan\phi = \frac{\Lambda_2}{\Lambda_1}$$

In principle, ϕ can be an arbitrary angle. However, in order to obtain the desired coupling coefficients (will be discussed later), ϕ should be a small angle and we will set $\Lambda_2 / \Lambda_1 = 1/3$, which gives $\phi \approx 18^\circ$. Shown in Fig. 7.26 is the primitive cell for such a PCDFB lattice, where the circular lattice element (shaded region) with a radius of r is used. In the reciprocal space, there are four primary reciprocal vectors denoted by $\mathbf{G}_1$, $\mathbf{G}_{-1}$, $\mathbf{G}_2$ and $\mathbf{G}_{-2}$, which stand for the major light propagation directions inside the PCDFB cavity.



Real space

Reciprocal space

Fig. 7.26. Sketch of the primitive cell for a PCDFB pattern in both real and reciprocal space. [Reproduced with permission from Proceedings of the SPIE Vol. 6900, Y. Bai, P. Sung, S.R. Darvish, W. Zhang, A. Evans, S. Slivken, and M. Razeghi, "Electrically pumped photonic crystal distributed feedback quantum cascade lasers," fig. 2, pg. 69000A-3, Copyright 2008 SPIE.]

7.8.2. Coupling coefficients

Analogous to the case of 1D-DFB, where the backward traveling light is coupled to the forward traveling light via a coupling coefficient κ (in cm^{-1}), there exit three coupling coefficients for the PCDFB pattern of Fig. 7.26. As mentioned before, one of the four primary reciprocal vectors needs to be aligned with the facet normal, which is the direction of the forward traveling light. So the coupling of the other three vectors to this vector will give three coupling coefficients:

Eq. (7.41) $$\kappa_i = \frac{2\pi\Delta n}{\lambda}\frac{1}{\Lambda_1\Lambda_2}\iint_s ds \exp\left(-i\mathbf{G}_{\kappa i}\cdot\mathbf{r}\right)$$

where i reads 1, 2, and 3, with:

Eq. (7.42) $$\mathbf{G}_{\kappa 1} = \mathbf{G}_{-1} - \mathbf{G}_1$$
Eq. (7.43) $$\mathbf{G}_{\kappa 2} = \mathbf{G}_2 - \mathbf{G}_1$$
Eq. (7.44) $$\mathbf{G}_{\kappa 3} = \mathbf{G}_{-2} - \mathbf{G}_1$$

The integration area s spans the lattice element and Δn is the refraction index contrast for the grating layer, shall InP be replaced by GaInAs. We can change Δn by changing the thickness and/or the position of the grating layer and this affects all three coupling coefficients in the same manner. However, if we change the shape and/or size of the lattice element, we can tune the relative ratio among the three coupling coefficients. For this purpose, let us define the pattern coupling coefficients:

Eq. (7.45) $$\kappa_i^p = \frac{1}{\Lambda_1\Lambda_2}\iint_s ds \exp\left(-i\mathbf{G}_{\kappa i}\cdot\mathbf{r}\right)$$

With $\Lambda_2/\Lambda_1 = 1/3$, we have the behavior of the pattern coupling coefficients as a function of the hole size factor plotted in Fig. 7.27, where the hole size factor β is defined as $\beta = 2r/\Lambda_2$.

Note that the pattern coupling coefficients are unitless and we need the value of Δn and λ to calculate the coupling coefficients in cm^{-1}. In reality, we set the targeting coupling coefficients before hand and changing Δn and β to reach the target.

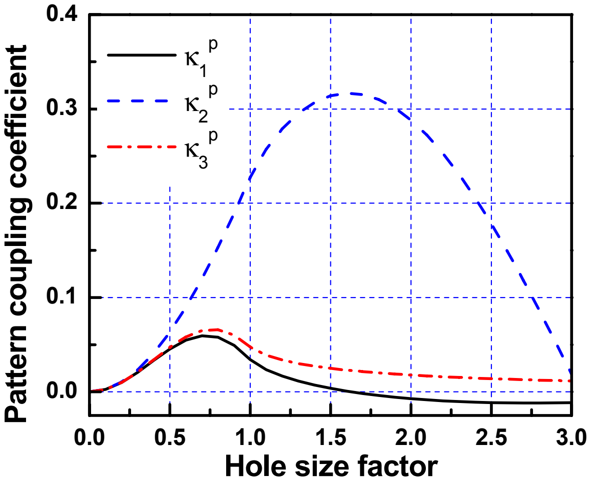


Fig. 7.27. Calculated pattern coupling coefficients as a function of the hole size factor.

7.8.3. Testing results

The fabrication of a photonic crystal distributed feedback (PCDFB) QCL starts with a wafer that has only the QCL core and a grating layer. The PCDFB pattern is defined in the grating layer by the electron beam lithography and dry etching. Then, an epitaxial regrowth is performed for the top cladding to finish the structure, as depicted in Fig. 7.25.

In general, excessive processing results in an efficiency penalty. As such, the slope efficiency of a PCDFB QCL is smaller than that of the Fabry-Perot counterpart with the same ridge width and cavity length. About one half of the slope efficiency of the Fabry-Perot QCL has been achieved for the PCDFB QCL. Further improvement is expected for better fabrication. In spite of the reduced slope efficiency, peak power as high as tens of watts is still achievable thanks to the big volume of the emitting core.

While the output power scales with the emitting volume, the essence of a PCDFB QCL lies in the spatial and spectral beam quality. Shown in Fig. 7.28 is the comparison of the far field and spectrum between a PCDFB QCL and a Fabry Perot QCL. The ridge width (100 μm) and cavity length (3 mm) are the same for both devices. It is clear that the Fabry Perot device is not single mode in either spatially or spectrally, but the PCDFB device is single mode both spatially and spectrally.

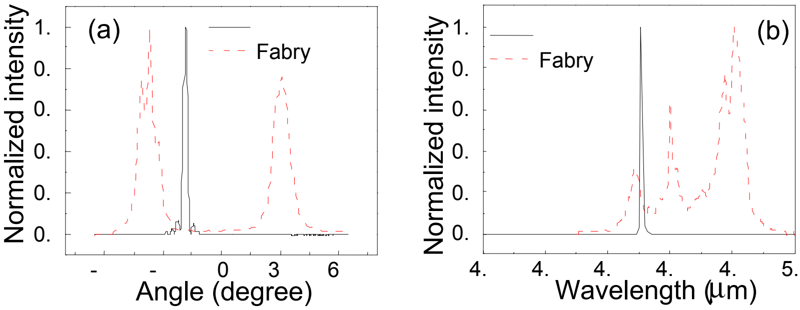


Fig. 7.28. Far field (a) and spectrum (b) comparison between a PCDFB QCL and a Fabry Perot QCL.

In particular, we can check whether the PCDFB far field is diffraction limited or not. The diffraction theory suggests that the full-width at half-maximum (FWHM) of the far field emission in the lateral direction has a minimum value set by the wavelength and the ridge width. This minimum FWHM can be obtained only when the optical field is uniform across the facet, in which case the laser far field restores the diffraction pattern of a plain wave illuminating a slit with the same width as the laser ridge width.

The Frauenhofer diffraction equation for a plain wave illuminating a slit reads:

Eq. (7.46) $$F(\theta) = \cos\theta \int_{x=-a/2}^{x=a/2} \exp[if(x)]\exp(-ikx\sin\theta)dx$$

where θ is the diffraction angle and $f(x)$ is the phase function of the plain wave. The width of the slit is denoted by a and the wavelength is contained in the wave vector $k = 2\pi/\lambda$. The phase function of the plain wave is linear across the facet:

Eq. (7.47) $$f(x) = kx\sin\delta$$

with δ denoting the illumination angle.

The comparison of the PCDFB far field with the Fraunhofer diffraction simulation is shown in Fig. 7.29. Despite an off normal emission, the close resemblance of these two curves clearly suggests that the PCDFB far field is indeed diffraction limited.

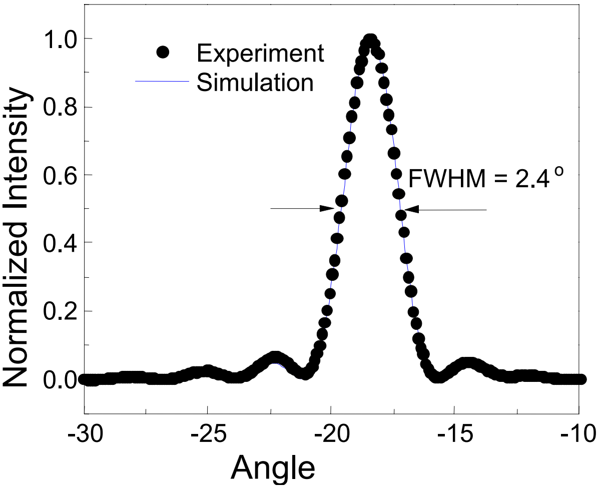


Fig. 7.29. Comparison of the experimentally measured PCDFB far field with the diffraction simulation of a plain wave illuminating a slit.

7.9. Quantum cascade lasers at different wavelengths

As pointed out previously, the wavelength of a QCL is determined by the energy difference of the upper laser level and the lower laser level. By choosing a proper material system and engineering the core heterostructure accordingly, a wide range of wavelengths is technically accessible. On the other hand, real world applications desire emitters in the mid-infrared spectral region and the terahertz spectral region, which facilities the development of mid-infrared QCLs and terahertz QCLs [Kohler et al. 2002]. At the moment, with room temperature CW operation of multi-watt optical power, mid-infrared QCLs perform significantly better than terahertz QCLs, which are technically more difficult.

7.9.1. Short wavelength quantum cascade lasers (<4 μm)

Due to the limited conduction band offset for the $Ga_xIn_{1-x}As/Al_yIn_{1-y}As/InP$ system, it becomes more and more difficult to extend the lower wavelength boundary for QCLs based on this material system. However, with the strain-balanced technique, room temperature continuious wave (CW) operation down to 3.8 μm has been demonstrated. The limitations for short wavelength QCLs are the following [Razeghi et al. 2005]:

- As the upper laser level is pushed toward the conduction band of the barrier, thermally activated carrier leakage into the continuum becomes the major factor that hinders high temperature operation.
- The intervalley scattering of electrons from the Γ valley into X and/or L valleys becomes more likely and may diminish the gain and population inversion of the emitter.
- A short wavelength QCL generally requires a higher voltage drop than longer wavelength devices due to the larger photon energy. This increases the power density of the core for a given current density. Therefore the thermal management for high temperature CW operation is more difficult.

7.9.2. Mid wavelength quantum cascade lasers (4–9 μm)

Mid-infrared QCLs made from the GaInAs/AlInAs/InP system are best suited for emission wavelengths between 4 μm and 9 μm. In this wavelength range, both the electron confinement and optical losses are more manageable for efficient operations. In fact, most of the pioneering works on QCLs were done in this wavelength range and most of the discussions in this chapter are based on middle wavelength QCLs.

7.9.3. Long wavelength quantum cascade lasers (>9 μm)

Unlike in the short wavelength side, where a lower boundary exists due to the finite conduction band offset, there is no boundary in the long wavelength side. In fact, the terahertz QCLs of hundreds of micrometers in wavelength, have been demonstrated. The core and waveguide design for the terahertz QCL are very different from that of the mid-infrared QCL and they are usually treated as an independent subject. For a mid-infrared QCL with a wavelength greater than 9 μm but less than 15 μm, the following considerations need to be addressed [Slivken et al. 2007]:

- The free carrier absorption, which is the main source of waveguide loss, increases roughly as the square of the wavelength. This makes it necessary to use lower doping and thicker layers to maintain low absorption loss.
- A thicker QCL core is needed to maintain the confinement factor for a longer wavelength. This increases the operating voltage and epitaxial thickness.
- The thermal resistance of the device increases as the core gets thicker.

7.10. Summary

In this chapter, we discuss the operation principles for the quantum cascade lasers (QCLs) based on a rate equation model. Important concepts, such as the intersubband transition, cascading, and polar optical phonon resonance, are explained in detail. The structure of a real QCL is then analyzed along with the technologies involved in making a QCL.

The performance of a QCL is discussed in terms of the power-current-voltage characteristics, temperature dependent characteristics, and most importantly the wall plug efficiency, which is an overall figure of merit for a QCL. Several established methods of increasing the wall plug efficiency are then discussed, followed by the discussion for power scaling.

The principle of the photonic crystal distributed feedback (PCDFB) is discussed showing the improvements in both the laser far field and spectrum characteristics.

Finally, the QCL's capability of covering a wide range of emitting wavelength is briefly reviewed, together with the challenges of extending the emission to shorter and longer wavelengths.

References

Bai, Y., Darvish, S.R., Slivken, S., Sung, P., Nguyen, J., Evans, A., Zhang, W., and Razeghi, M., "Electrically pumped photonic crystal distributed feedback quantum cascade lasers," *Applied Physics Letters* 91, p. 141123, 2007.

Bai, Y., Darvish, S.R., Slivken, S., Zhang, W., Evans, A., Nguyen, J., and Razeghi, M., "Room temperature continuous wave operation of quantum cascade lasers with watt-level optical power," *Applied Physics Letters* 92, p. 101105, 2008a.

Bai, Y., Slivken, S., Darvish, S.R., and Razeghi, M., "Room temperature continuous wave operation of quantum cascade lasers with 12.5% wall plug efficiency," *Applied Physics Letters* 93, p. 021103, 2008b.

Bartolo, R.E., Bewley, W.W., Vurgaftman, I., Felix, C.L., Meyer, J.R., and Yang, M.J., "Mid-infrared angled-grating distributed feedback laser," *Applied Physics Letters* 76, pp. 3164-3166, 2000.

Darvish, S.R., Slivken, S., Evans, A., Yu, J.S., and Razeghi, M., "Room-temperature, high-power, and continuous-wave operation of distributed-feedback quantum-cascade lasers at lambda similar to 9.6 μm," *Applied Physics Letters* 88, p. 201114, 2006a.

Darvish, S.R., Zhang, W., Evans, A., Yu, J.S., Slivken, S., and Razeghi, M., "High-power, continuous-wave operation of distributed-feedback quantum-cascade lasers at lambda similar to 7.8 μm," *Applied Physics Letters* 89, p. 251119, 2006b.

Evans, A., Yu, J.S., Slivken, S., and Razeghi, M., "Continuous-wave operation of $\lambda \sim 4.8$ μm quantum-cascade lasers at room temperature," *Applied Physics Letters* 85, pp. 2166-2168, 2004.

Faist, J., Capasso, F., Sivco, D.L., Sirtori, C., Hutchinson, A.L., and Cho, A.Y., "Quantum cascade laser," *Science* 264, pp. 553-556, 1994.

Faist, J., Gmachl, C., Capasso, F., Sirtori, C., Sivco, D.L., Baillargeon, J.N., and Cho, A.Y., "Distributed feedback quantum cascade lasers," *Applied Physics Letters* 70, pp. 2670-2672, 1997.

Kazarinov, R.F. and Suris, R.A., "Possibility of Amplication of Electromagnetic Waves in a Semiconductor with a Superlattice," *Soviet Physics Semiconductors-Ussr* 5, p. 707, 1971.

Kohler, R., Tredicucci, A., Beltram, F., Beere, H.E., Linfield, E.H., Davies, A.G., Ritchie, D.A., Iotti, R.C., and Rossi, F., "Terahertz semiconductor-heterostructure laser," *Nature* 417, pp. 156-159, 2002.

Nguyen, J., Yu, J.S., Evans, A., Slivken, S., and Razeghi, M., "Optical coatings by ion-beam sputtering deposition for long-wave infrared quantum cascade lasers," *Applied Physics Letters* 89, p. 111113, 2006.

Razeghi, M., Evans, A., Slivken, S., and Yu, J.S., "High power CW quantum cascade lasers: How short can we go?," *Proceedings of SPIE* 5738, pp. 1-12, 2005.

Slivken, S., Evans, A., Zhang, W., and Razeghi, M., "High-power, continuous-operation intersubband laser for wavelengths greater than 10 µm," *Applied Physics Letters* 90, p. 151115, 2007.

Slivken, S., Evans, A., Nguyen, J., Bai, Y., Sung, P., Darvish, S.R., Zhang, W., and Razeghi, M., "Overview of quantum cascade laser research at the center for quantum devices," *Proceedings of SPIE* 6900, pp. 69000B1-8, 2008.

Smet, J.H., Fonstad, C.G., and Hu, Q., "Intrawell and interwell intersubband transitions in multiple quantum wells for far-infrared sources," *Journal of Applied Physics* 79, pp. 9305-9320, 1996.

Vurgaftman, I. and Meyer, J.R., "Photonic-crystal distributed-feedback quantum cascade lasers," *IEEE Journal of Quantum Electronics* 38, pp. 592-602, 2002.

Yu, J.S., Slivken, S., Darvish, S.R., Evans, A., Gokden, B., and Razeghi, M., "High-power, room-temperature, and continuous-wave operation of distributed-feedback quantum-cascade lasers at lambda similar to 4.8 µm," *Applied Physics Letters* 87, p. 041104, 2005.

Yu, J.S., Evans, A., Slivken, S., Darvish, S.R., and Razeghi, M., "Temperature dependent characteristics of $\lambda \sim 3.8$ µm room-temperature continuous-wave quantum-cascade lasers," *Applied Physics Letters* 88, p. 251118, 2006.

Further reading

Agrawal, G. and Dutta, N., *Semiconductor Lasers*, Van Nostrand Reinhold, New York, 1993.

Chuang, S.L., *Physics of Optoelectronic Devices*, John Wiley & Sons, New York, 1995.

Iga, K., *Fundamentals of Laser Optics*, Plenum Press, New York, 1994.

O'shea, D., *Introduction to Lasers and Their Applications*, Addison-Wesley, Reading, MA, 1978.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Razeghi, M., *The MOCVD Challenge Volume 2: A Survey of GaInAsP-GaAs for Photonic and Electronic Device Applications*, Institute of Physics, Bristol, UK, pp. 21-29, 1995.

Razeghi, M., "Optoelectronic Devices Based on III-V Compound Semiconductors Which Have Made a Major Scientific and Technological Impact in the Past 20 Years," *IEEE Journal of Selected Topics in Quantum Electronics*, 2000.

Razeghi, M., "Kinetics of Quantum States in Quantum Cascade Lasers: Device Design Principles and Fabrication," *Microelectronics Journal* 30, pp. 1019-1029, 1999.

Siegman, A.E., *Lasers*, University Science Book, Mill Valley, Calif., 1986.

Silfvast, W.T., *Laser Fundamentals*, Cambridge University Press, New York, 1996.

Streetman, B.G., *Solid States Electronic Devices,* Prentice-Hall, Englewood Cliffs, NJ, 1990.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Problems

1. A model to describe the dynamic behavior of the carrier concentration must fulfill the current conservation for the steady state solution, which is a fundamental constrain. Prove that the current conservation is built-in in the rate equation model in this chapter. (Hint: find the expression for the outward current term and see if it equals the inward current term, for the cases of both below and above threshold.)

2. Calculate the net strain of the following system (in ppm): $Al_{0.65}In_{0.35}As$ (3 nm) / $Ga_{0.33}In_{0.67}As$ (5 nm) / $Al_{0.65}In_{0.35}As$ (3 nm), pseudomorphically grown on InP substrate. If the absolute value of the net strain is less than 1000 ppm, we call it strain balanced. Is it strain balanced? If not, find at least one solution to approach a strain-balanced structure. Use the parameters in the following table for the calculation. (Hint: use Vegard's law for the elastic constants of ternaries.)

	a(Å)	c_{11} (GPa)	c_{12} (GPa)
AlAs	5.6622	1250	534
GaAs	5.6532	1221	566
InAs	6.0584	833	453
InP	5.8687	-	-

3. The measure for the quality of the thermal packaging is the thermal resistance R_{th}, which is defined by $R_{th} = \frac{\Delta T}{P_e}$, where an input electrical

power of P_e results a temperature rising of ΔT for the laser core. Calculate the thermal resistance of the device shown in Fig. 7.16 at the threshold of the CW operation.

4. Examine the expression for the electrical efficiency η_e (Eq. (7.34)). It is obvious that $\eta_e = 0$ when $J = J_{th}$, and $\eta_e \to 0$ when $J \to +\infty$. This means that there exists a critical value J_{peak}, at which η_e reaches the maximum of $\eta_e^{\max}$. Find J_{peak} and $\eta_e^{\max}$.

5. The graphic method of extracting the internal quantum efficiency η_i and the waveguide loss α_w by fitting the data of many devices with different mirror losses is statistically more accurate. However, this is not necessary if we can accurately measure the slope efficiency. In principle, we only need two equations to determine two unknown parameters. Given the external quantum efficiency of η_{ext1} and η_{ext2}, and the mirror loss of α_{m1} and α_{m2} for two QCLs, construct two equations and find the expression for η_i and α_w. Now we have two devices from the same double channel processed QCL wafer. The wavelength is 4.6 µm and the laser core consists of 30 QCL stages. The cavity length of both devices is 3 mm. One of them is uncoated and the other is HR coated. The slope efficiency is measured to be 1.72 W/A for the uncoated device and 2.86 W/A for the HR coated device, with both values obtained from the front facet only. Calculate the internal quantum efficiency and the waveguide loss for this double channel processed QCL wafer.

8. Photodetectors: General Concepts

8.1. Introduction
8.2. Electromagnetic radiation
8.3. Photodetector parameters
  8.3.1. Responsivity
  8.3.2. Noise in photodetectors
  8.3.3. Noise mechanisms
  8.3.4. Detectivity
  8.3.5. Detectivity limits and BLIP
  8.3.6. Frequency response
8.4. Thermal detectors
8.5. Summary

8.1. Introduction

A detector can be defined as a device that converts one type of signal into another as illustrated in Fig. 8.1. Various forms of input signal can be entered into the detector, which then generates the measurable output signal, such as an electrical current or voltage. There exist many different types of detectors depending on the objects or physical properties that they sense. The input signal can be mechanical vibrations, electromagnetic radiation, small particles, and other physical phenomena. Smoke detectors can sense the soot particulates caused by fire and seismometers sense the mechanical vibrations caused by the earth. The human body has various types of detectors: the eyes can sense electromagnetic radiation in the visible range, the ears detect sound from pressure variations through a medium such as atmospheric air or water, the tongue senses various types of chemicals, and the skin can detect temperature and pressure. Our natural sensory skills have been augmented through the development of advanced instruments such as the microscope and the thermometer, that were made possible thanks to the development of technology. Furthermore technology has made it possible for humans to detect things that could not be naturally sensed by the human body. For example, we can observe the infrared (IR) light emitted from

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_8,
© Springer Science+Business Media, LLC 2010

warm objects and the ultraviolet (UV) light from hot objects with the help of photodetectors which will be the focus of the this and the next Chapters.

Input signal
Detector
Output signal

Fig. 8.1. Concept of a detector. The input signal usually has the form of electromagnetic radiation and the output signal is often an electrical signal. The detector is a device which converts one type of signal into another which can then be processed.

Human eyes respond only to visible light, from violet to red. However, the light spectrum is much broader and includes radiation beyond violet (e.g. ultraviolet, gamma rays) and red (e.g. infrared, microwaves). If the temperature of the object is larger than 6,000 K, it will emit predominantly in the ultraviolet. However, colder objects (<2,000 K) emit predominantly in the infrared. Most of the objects on earth emit IR light, and by choosing the correct materials and growth and fabrication techniques, photodetectors can be designed to sense light in this wavelength range. Using infrared photodetectors, we can obtain information on the objects emitting this radiation to determine their geometry, temperature, surface quality, and chemical content. We can also get information on the atmosphere through which the IR light is propagated.

Due to the fact that some wavelengths of infrared light are transmitted with little loss within the Earth's atmosphere, the IR spectrum offers some attractive advantages for photodetection purposes. Because of this and other advantages, IR photodetectors have been in active development over the last several decades and found numerous applications, such as night vision, missile guidance, and range finders. As the cost of these IR photodetectors has decreased, they have become more available for civilian and industrial applications where they are used in hazardous gas sensing, security systems, thermal imaging for medical purposes, hot spot monitoring and optical communications. Specialized infrared imagers have recently been used to detect malignant cancers and have acted as collision and ranging sensors in automobiles. Due to their prominence in commercial and military applications, in this and the next Chapters we will focus on photodetectors designed for the infrared regime.

Regardless of sensing wavelength, photodetectors are usually integrated into a system that generates a signal which can be easily recognized and interpreted by humans. A few elements of such a system are shown in a block diagram form in Fig. 8.2. The system may be designed to detect the target, to track it as it moves, or to measure its temperature. If the radiation from the target passes through any portion of the earth's atmosphere, it will be *attenuated* because the atmosphere is not perfectly transparent. The *optical receiver* collects some of the radiation from the target and delivers it to a detector which converts it into an electrical signal. Before reaching the *detector*, the radiation may pass through an *optical modulator* where it is coded with information concerning the direction to the target or information destined to assist in the discrimination of the target from unwanted details in the background. Since some detectors must be cooled, one of the system elements may be a *cooler*. The electrical signal from the detector then passes through a *processor* where it is amplified and the coded target information is extracted. The final step is the use of this information to automatically control some process or to display the information for interpretation by a human observer.

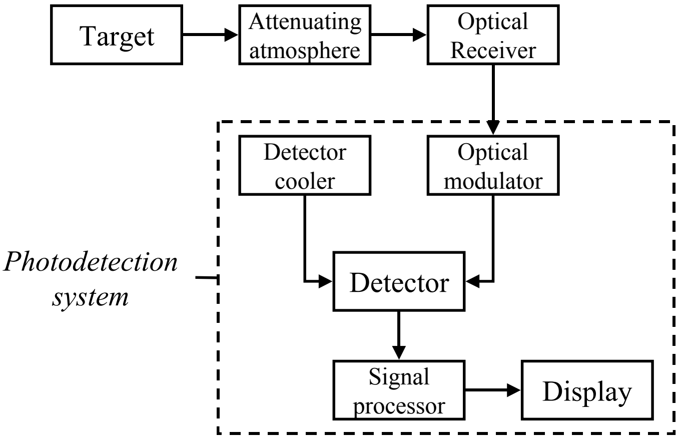


Fig. 8.2. The major elements constituting a photodetection system.

In this Chapter, we will first review the fundamental concepts of electromagnetic radiation emitted by a body, which will allow us to better understand the principles of photodetection. Next, we will describe the theory of operation of photodetectors and introduce the important parameters that characterize and compare their performance. Most photodetectors can be divided in two types: thermal detectors and photon detectors, and the difference lies in the detection mechanism. We will explore thermal detectors here and discuss the operation and some specific examples of photon detectors in the following Chapter.

8.2. Electromagnetic radiation

The schematic in Fig. 8.3 shows various types of electromagnetic radiation along with their associated wavelength and frequency ranges. The borders between visible, infrared, far-infrared, and millimeter waves are not absolute, and have been introduced primarily for convenience. For example, visible light is that portion of the spectrum to which the human eye is sensitive, and a statement such as "infrared extends from 0.7 μm to 1000 μm" is only a convention. Typically, IR radiation does not penetrate metals unless these are very thin, but passes through many crystalline, plastic, and gaseous materials-including the earth's atmosphere. There does not exist a detector that can detect all types of radiation with the same sensitivity. Thus, a photodetector has to be designed to operate within a specific spectral bandwidth.

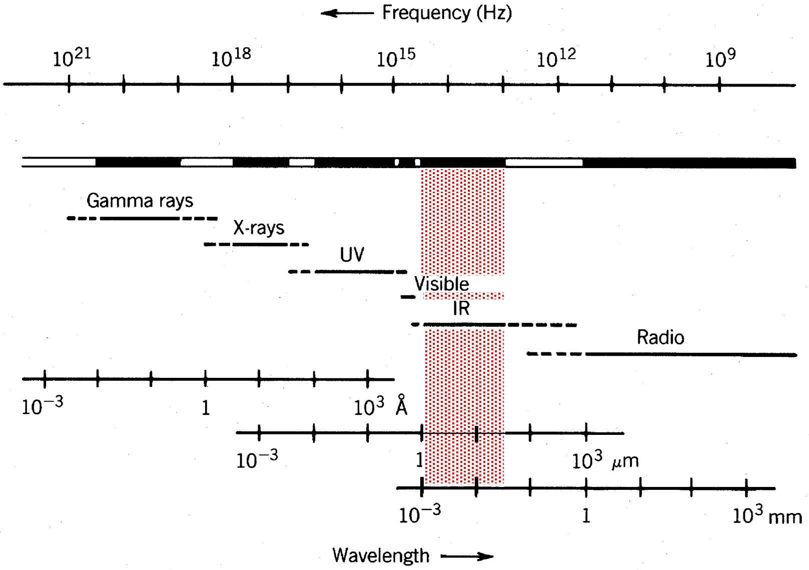


Fig. 8.3. The electromagnetic spectrum. The major spectral regions of interest here are shown with their limits in terms of frequency and wavelength, including the ultraviolet (UV), visible and infrared (IR).

A source of electromagnetic radiation generally emits over a broad range of wavelengths, and some wavelengths are emitted with more power than others. For example, different bodies emit radiation differently depending on their surface properties, due to varying emissivities. To have an absolute scale for proper comparison, we typically use a blackbody which is a perfect absorber of all radiant energy and which is also a perfect emitter of electromagnetic radiation. The intensity of the blackbody radiation depends on the wavelength of emitted light. This dependence is called the spectral distribution of intensity and is a function of the blackbody's absolute temperature.

An analytical expression of this spectral intensity distribution is shown in Eq. (8.1) and was determined by Max Planck in 1901. His theory assumed that the energy carried by an electromagnetic radiation is composed of discrete or quantized energy packages proportional to the frequency considered. The idea answered so many unsolved physics problems that Planck's hypothesis quickly became the basis for modern quantum theory. Planck's law, for which he received the Nobel Prize, quantifies the spectral radiation M from a blackbody as

Eq. (8.1) $$M(\lambda)=\frac{2\pi hc^2}{\lambda^5}\frac{1}{\exp\left(\frac{ch}{k_bT\lambda}\right)-1}\ (\mathrm{W\cdot m^{-2}\cdot m^{-1}})$$

where λ is the wavelength in meters, c is the velocity of light in vacuum in $\mathrm{m\cdot s^{-1}}$, T is the absolute temperature in K, h is Planck's constant and k_b is Boltzmann's constant. This relation expresses thermal radiation as a function of wavelength and temperature for all wavelengths. The peak wavelength of the ideal blackbody emission is described by Wien's law:

Eq. (8.2) $$\lambda_{Peak}(T_B)=\frac{hc}{4.965k_bT_B}=\frac{2.898\times10^{-3}K\cdot m}{T_B}$$

When considering the entire spectral power R of an ideal blackbody (integrating Planck's Law over all wavelengths), the result is the Stephan-Boltzmann law for total blackbody emittance:

Eq. (8.3) $$\int_0^\infty M(\lambda,T)d\lambda=R(\lambda,T)=\sigma T^4$$

where σ is the Stephan-Boltzmann constant $5.67\times10^{-8}\ \mathrm{W\cdot m^{-2}\cdot K^{-4}}$.

8.3. Photodetector parameters

As mentioned earlier, photodetectors are devices which sense light (as an input signal) and generate a measurable output signal in the form of an electrical current or voltage. The performance of these photodetectors can be quantified and compared using several parameters, which will be discussed in this section.

8.3.1. Responsivity

The responsivity of a photodetector is the ratio of its output electrical signal, either a current I_{out} or a voltage V_{out}, to the input optical signal expressed in terms of the incident optical power P_{in}. One can define a current responsivity and a voltage responsivity using respectively:

Eq. (8.4)
$$R_i = \frac{I_{out}}{P_{in}} \text{ and } R_v = \frac{V_{out}}{P_{in}}$$

The current responsivity R_i is expressed in terms of A/W, while the voltage responsivity R_v is expressed in units of V/W. The output signals I_{out} and V_{out} in Eq. (8.4) can be expressed in detail for the case of specific light detection mechanisms of the detector in question. The output current and voltage are often called photocurrent (I_{ph}) and photovoltage (V_{ph}) as they arise in the presence of light. The incident input power P_{in} on the photodetector can be expressed as:

Eq. (8.5)
$$P_{in} = A\Phi_{ph}\frac{hc}{\lambda}$$

where A is the area of the detector, Φ_{ph} is the incident photon flux density expressed in units of photons$\cdot$m$^{-2}\cdot$s^{-1}, h is Planck's constant, c is the velocity of light in vacuum, and λ is the wavelength of the incident light.

8.3.2. Noise in photodetectors

As the output signal of a photodetector is an electrical signal, it is not a strictly stable quantity over time, but fluctuates due to electrical noise. Electrical noise is a random variable that results from stochastic or random processes associated with various particles, i.e. discrete, nature of electrons, phonons, and photons and their interactions. For example, the instantaneous voltage signal $V(t)$ is shown schematically in Fig. 8.4. A quantitative measure of the noise in the voltage signal is then its root-mean-square and is given by:

Eq. (8.6)
$$I_{noise} = \sqrt{\frac{1}{T}\int_0^T \left(I(t) - \langle I \rangle\right)^2 dt}$$

This relation expresses the average value of the square of the fluctuation of the variable $I(t)$ around its average $\langle I \rangle$ over a long period of time T.

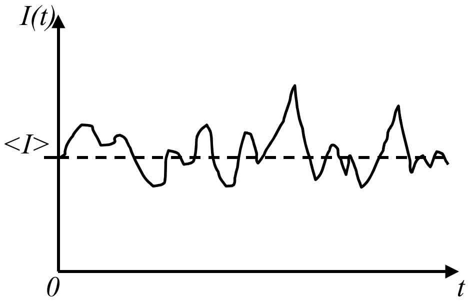


Fig. 8.4. Illustration of the instantaneous current signal I(t) exhibiting electrical noise about its average value ⟨I⟩.

It is also useful to define a power signal-to-noise ratio as:

Eq. (8.7) $$\frac{S}{N}=\frac{I_{ph}^{2}}{\left\langle I_{noise}^{2}\right\rangle}$$

where $\langle\ldots\rangle$ denotes the statistical average of a random variable, I_{ph} is the instantaneous photocurrent and I_{noise} is the instantaneous noise current. We talk about a power ratio because the expression in Eq. (8.7) involves the squares of the electrical current. The effect of the signal-to-noise ratio is illustrated in Fig. 8.5.

The noise current I_{noise} is usually a function of the frequency considered, and the noise power I_{noise}^{2} depends on the frequency bandwidth Δf over which the statistical average is measured. For example, if the output signal is centered at a frequency of 200 Hz, eliminating all electrical signal outputs with significant components above 250 Hz and below 150 Hz would decrease the noise power in the output because the remaining 100 Hz electrical bandwidth contains less noise than the entire frequency band in the case of white noise. This is why, in order to compare different noise mechanisms in different photodetectors, with different frequency bandwidths, it is convenient to use the concept of noise spectral density, expressed in terms of $\mathrm{A}\cdot\mathrm{Hz}^{-1/2}$, which consists of normalizing the noise power to the frequency bandwidth considered:

Eq. (8.8) $$\text{noise spectral density}=\frac{\sqrt{\left\langle I_{noise}^{2}\right\rangle}}{\sqrt{\Delta f}}$$

A similar expression can be obtained using the voltage noise V_{noise}.

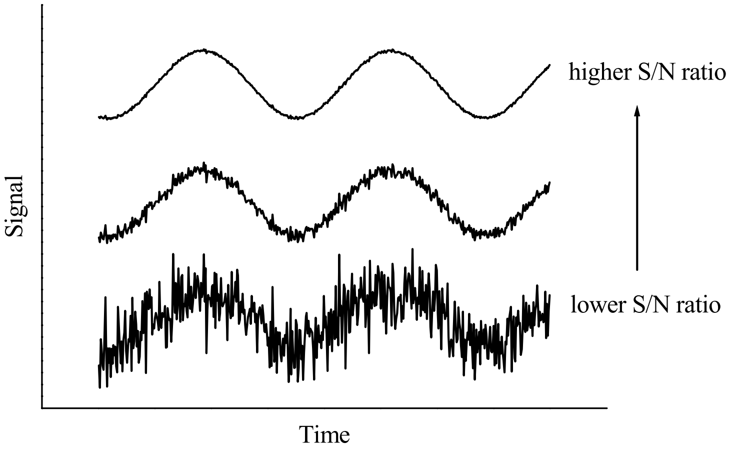


Fig. 8.5. Detector output with varying signal-to-noise ratios. When the S/N ratio is high, the signal is clear but when the S/N ratio is low, the random modulation signal (noise) is superimposed which reduces the accuracy of the actual signal.

A useful concept in the evaluation of the electrical noise is to consider a noise equivalent circuit that consists of a small signal electrical circuit model of the device including the noise source. This equivalent circuit makes use of the differential resistance R_{diff} of the device when a bias voltage $V = V_b$ is applied:

Eq. (8.9) $$R_{diff} = \left(\frac{\partial V}{\partial I} \right)_{|V=V_b}$$

where V and I are the voltage and current of the device, respectively. The noise equivalent circuits are shown in Fig. 8.6.

These noise equivalent circuits can only be used when the noise in the devices is a white noise, i.e. does not depend on the frequency. These noise equivalent circuits can be used like any small signal circuits when several contributions to the total noise are considered, with the exception that the noise powers need to be added instead of summing or subtracting noise voltages and noise currents, e.g.:

Eq. (8.10) $$\left\langle I_{noise}^2 \right\rangle_{total} = \left\langle I_{noise}^2 \right\rangle_1 + \left\langle I_{noise}^2 \right\rangle_2$$

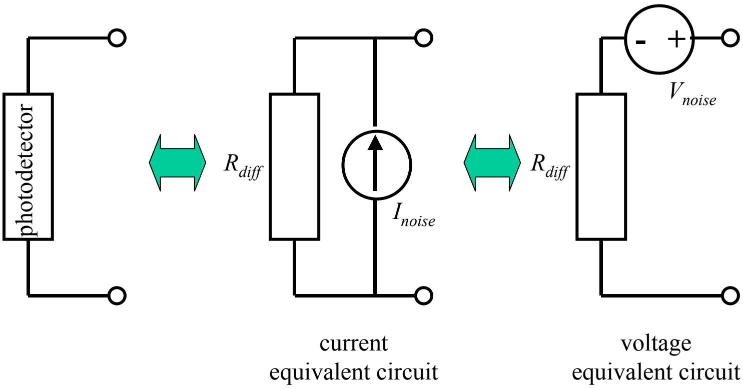


Fig. 8.6. Current and voltage equivalent circuits of a photodetector, including the sources of noise I_{noise} and V_{noise}.

8.3.3. Noise mechanisms

There exist several contributions to the noise in photodetectors that will be briefly described in this sub-section.

Johnson noise.

Thermal or Johnson noise occurs in all electrically conductive materials regardless of the conduction mechanism and results from the random motion of thermally-activated charge carriers through the conductor. The total noise current is proportional to the sum of the carrier movement occurring within a short time frame. Johnson noise is named after the scientist who experimentally investigated it [Johnson 1928]. The mean-square voltage of such noise can be calculated from:

Eq. (8.11) $$\left\langle V_{noise}^2 \right\rangle_{Johnson} = 4k_b TR\Delta f$$

where k_b is the Boltzmann constant, T is the temperature of the conductor, R is its electrical resistance, and Δf is the frequency bandwidth of the noise which is the frequency range in which the noise exists or is considered. This equation shows that the value of the noise in a given bandwidth is independent of the value of frequency, therefore the Johnson noise is a white noise. As a result, there is a constant maximum noise power P_{noise} from any resistor R at a given temperature T:

Eq. (8.12) $$\left(P_{noise}\right)_{Johnson\max} = \frac{1}{2}\frac{\left\langle V_{noise}^2 \right\rangle}{2R} = k_b T\Delta f$$

and the value of this power at room temperature is about 4 × 10^{-21} W.

Shot noise.

In 1918, Schottky showed that the random arrival of electrons on the collecting electrode of a vacuum tube was responsible for a so-called shot noise. At the origin of shot noise is the process of charge carriers being thermally or optically excited over a junction barrier. The current shot noise in a simple temperature limited vacuum diode is given by the following expression:

Eq. (8.13) $$\left\langle I_{noise}^2 \right\rangle_{shot} = 2qI_{DC}\Delta f$$

where q is the elementary charge and I_{DC} is the DC current bias of the vacuum diode. This equation shows that shot noise is also independent of the frequency and is a white noise. However, this is only valid if the inverse of the frequency of operation, $1/f$, is much larger than the traveling time of the electron in the device.

The shot noise in a *p*-*n* junction diode can be estimated by the following equation for the low frequency region:

Eq. (8.14) $$\left\langle I_{noise}^2 \right\rangle_{shot} = 2q\left(I_D + 2I_0\right)\Delta f$$

where $I_D = I_0\left(e^{\frac{qV_b}{k_bT}} - 1\right)$ and I_0 is the saturation current of the diode. This relation is equivalent to Eq. (8.11) when no bias is applied ($V_b = 0$), and to Eq. (8.13) for a high enough current bias. At large negative bias the noise in Eq. (8.14) approaches that of Eq. (8.13).

1/f noise.

There also exists an important type of noise which has a power spectrum inversely proportional to the frequency of operation f. This noise mechanism often dominates other mechanisms at low frequencies. This is a process dependent noise in the sense that it can be affected by the contact type and preparation, as well as the surface preparation and passivation. For example, carrier trapping and re-emission to and from defects at surfaces and contacts may contribute to this $1/f$ noise. It should be noted that a variety of names have been used for this type of noise in the literature such as excess noise, modulation noise, contact noise, and flicker noise. The power spectrum of this noise is given by:

Eq. (8.15) $$P_{noise}(f) = \gamma \frac{I_{DC}^{\alpha}}{f^{\beta}}$$

where γ is a constant, I_{DC} is the DC current bias of the device, f is the frequency of operation, α and β are exponents characteristic of the particular device considered. The value of α is usually near 2, while the value of β ranges from 0.8 to 1.5.

Generation-recombination noise.

The generation and recombination of charge carriers in semiconductors are random processes as they are associated with the creation and annihilation of electron-hole pairs in the material. The number of free carriers during a given period of time is therefore not constant but fluctuates in a random manner too. This results in a random change of the voltage of the device which is called the generation-recombination or G-R noise. For a near intrinsic semiconductor with a moderate bias, this noise can be expressed as [Long 1967]:

Eq. (8.16) $$(V_{noise})_{G-R} = \frac{2V_b}{(lwt)^{1/2}}\left(\frac{1+b}{bn+p}\right)\left[\frac{np}{n+p}\frac{\tau\Delta f}{1+\omega^2\tau^2}\right]^{1/2}$$

where V_b is the voltage bias, l is the length, w is the width, and t is the thickness of the device, b is the ratio of mobilities μ_e/μ_h, n and p are the electron and hole concentrations respectively, τ is the carrier lifetime in the semiconductor, ω is the angular frequency of operation, and Δf is the frequency bandwidth within which the noise is measured.

Temperature noise.

If the conductivity of the device strongly depends on the temperature, any random temperature fluctuation would result in a so called temperature noise in the device. This is an important source of noise for all infrared thermal detectors as well as low noise preamplifiers. This is the reason why even uncooled infrared thermal detectors, i.e. which can operate without cooling, usually have a thermoelectric cooler to stabilize their temperature.

Photon noise.

This noise mechanism arises from the random arrival of photons onto the surface of the photodetector. Unlike the previously described noise mechanisms, photon noise is a source of noise which is extrinsic to the device. The equivalent power of this noise or noise-equivalent-power for a given detector with an area A and a quantum efficiency η is:

$$\text{Eq. (8.17)} \qquad \left(P_{NEP}\right)_{photon} = h\nu\left[\frac{2A\Phi_{bkg}\Delta f}{\eta}\right]^{1/2}$$

where Φ_{bkg} is the total background photon flux density reaching the detector, and υ is the frequency of the photon.

8.3.4. Detectivity

Although the responsivity of a photodetector gives a measure of the output signal of the detector for a given optical input signal, it does not give any information about the sensitivity of the device. The sensitivity of the detector can be defined as the minimum detectable optical input power that can be sensed with a signal-to-noise ratio of unity. This power is called the noise-equivalent-power (P_{NEP}) of the detector and is given by:

$$\text{Eq. (8.18)} \qquad P_{NEP} = \frac{I_{noise}}{R_i} \text{ or } P_{NEP} = \frac{V_{noise}}{R_v}$$

Jones suggested to define the detectivity of a detector as the inverse of this noise-equivalent-power [Jones 1953]:

$$\text{Eq. (8.19)} \qquad D = \frac{1}{P_{NEP}}$$

which is expressed in units of W^{-1}. This quantity is very useful when measuring the sensitivity of photodetectors. However, it is not a fair means of comparing the overall performance of different detectors because it neglects the effects of the detector area and frequency bandwidth. For example, photodetectors with different sizes and thus detection areas will have different noise-equivalent-powers. In addition, a detector with low electrical bandwidth can have higher detectivity than an otherwise identical detector with wider electrical bandwidth. This is despite the fact that higher bandwidth is desired for faster devices. To address these issues, Jones introduced the concept of specific detectivity, which is denoted $D*$ and is the detectivity of a photodetector with an area of 1 cm^2 and an electrical bandwidth of 1 Hz:

$$\text{Eq. (8.20)} \qquad D* = D\sqrt{A\Delta f} = \frac{\sqrt{A\Delta f}}{P_{NEP}}$$

where A is the area of the detector in cm^2. $D*$ is expressed in units of cm·Hz$^{1/2}$·W^{-1}. Since it is independent of the device dimensions and the electrical configuration used for the measurement, $D*$ is widely used to compare photodetectors with very different physical and operational characteristics and is often simply called detectivity.

One can easily express $D*$ in terms of the detector responsivity:

$$\text{Eq. (8.21)} \qquad D* = \frac{R_i}{I_{noise}}\sqrt{\frac{A}{\Delta f}} = \frac{R_v}{V_{noise}}\sqrt{\frac{A}{\Delta f}}$$

where Δf is the frequency bandwidth of the measurement setup, I_{noise} and V_{noise} are the total root-mean-square current and voltage noises of the detector in the given frequency bandwidth of Δf.

8.3.5. Detectivity limits and BLIP

In order to ascertain the maximum performance of a photodetector, it is important to understand the role of the background noise current, I_d. This current can arise from blackbody radiation absorbed by the detector from the environment with temperature T_B. Additionally background noise can be produced by the detector device itself in the form of dark currents, that is, noise contributions from conduction or leakage currents under no illumination. These two components define the upper detectivity limit photodetectors as shown in Fig. 8.7.

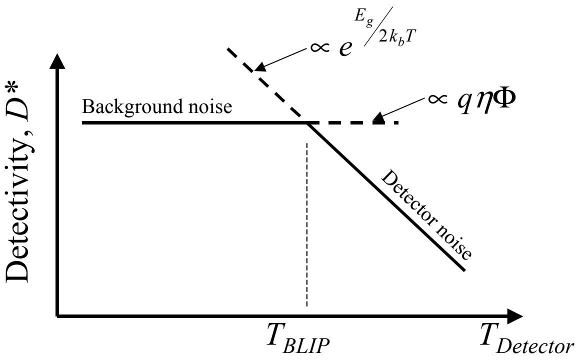


Fig. 8.7. Detectivity limit with respect to temperature for a photodetector.

As the detector temperature increases, the dark current increases exponentially as $e^{-E_g/2k_bT}$. Conversely, it is logical that the background noise stays constant with temperature and is proportional to the quantum efficiency of the device and the photon flux incident on the detector. The temperature at which the background noise begins to limit the device performance is known as the background limited infrared performance (BLIP) temperature. The BLIP threshold can be found from solving an equality between the constant background blackbody noise current and the device dark current:

Eq. (8.22) $$I_{sat}(T_{BLIP}) = I_B = q\int_{\lambda_1}^{\lambda_2} \eta(\lambda)\frac{d\Phi_B}{d\lambda}d\lambda$$

where η is the detector quantum efficiency (essentially the ratio between the absorbed and total incident photons) with respect to wavelength and $\frac{d\Phi_B}{d\lambda}$ is the photon flux over the wavelength range from λ_1 to λ_2. Using this formalism, one can also calculate the absolute detectivity limit for ideal photodiodes, for example:

Eq. (8.23) $$D^*_{BLIP,\max}(\lambda_0, T_B) = \frac{1}{h\nu}\frac{1}{2^{1/2}[\Phi_B(\lambda_0)]^{1/2}}$$

It should be noted that this limit should be modified by 1/sin(φ) when considering a real detection system with cryostat, and an acceptance angle of φ. A complete explanation of detectivity limits and BLIP can be found in [Rosencher and Vinter 2002].

8.3.6. Frequency response

When the optical input signal is periodic with a fixed amplitude and a frequency *f*, the amplitude of the detector electrical output signal is not necessarily a constant but may vary with the frequency as shown in Fig. 8.8. This phenomenon is usually due to the electrical resistive-capacitance or RC delay of the device in the case of the photon detectors, and the thermal RC delay of the thermal detectors. The frequency dependent responsivity can be approximated by:

Eq. (8.24) $$R_{i,v}(f) = R_{i0,v0}\frac{1}{\sqrt{1+\left(\frac{f}{f_c}\right)^2}}$$

where $R_{i0,v0}$ is the responsivity of the photodetector at very low frequencies, and:

Eq. (8.25) $$f_c = \frac{1}{2\pi RC}$$

where R is the electrical (or thermal) resistance and C is the electrical (or thermal) capacitance of the photon (or thermal) detector.

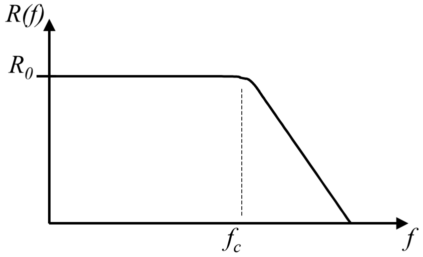


Fig. 8.8. Illustration of the frequency dependence of detector responsivity on a log-log scale."

8.4. Thermal detectors

In thermal detectors, the absorption of IR light leads to a change in the temperature of the detector, thus resulting in a change in resistance (bolometer) or electrical polarization (pyroelectric detectors). This change is then recorded by an electrical circuit. Because they can operate at room temperature, thermal detectors are mostly used whenever cooling systems are not possible and to minimize cost. It is however often necessary to use thermoelectric coolers in order to stabilize the temperature of the detectors, in an effort to minimize the Johnson noise. In general, thermal detectors have a large spectral bandwidth compared to their photon detector counterparts. This is because thermal-type devices typically absorb photons like a blackbody, or with a wide, flat response function with respect to wavelength. There exist several types of thermal detectors that will be briefly discussed here: bolometers, thermopiles, thermocouples, and pyroelectric detectors.

A bolometer is a thermal detector whose resistance depends on its temperature. Since the resistance of most semiconductors is a strong function of temperature, the resistance of a semiconductor chip can tell us how much radiant energy is falling on it. The Golay cell uses the expansion of a gas when heated to sense radiated power: the gas is contained in a chamber that is closed with a reflecting membrane. When the gas is warmed, the membrane distorts and deflects a beam of light that has been focused on it.

A thermopile typically consists of several thermocouple stacks in series. The principle of a thermocouple resides in the thermoelectric effect which yields an voltage proportional to the temperature difference between two dissimilar metal junctions. Therefore, an increase in the incident infrared radiation power causes an increase in the thermopile voltage.

When ferroelectric polar crystals are exposed to a change in temperature, their internal electric polarization changes, electrical charges accumulate and can be measured on opposite sides of the crystal. The capacitance of the detector material also changes and can be electrically measured. Pyroelectric detectors use this material property to detect IR radiation.

Among all these thermal detectors, bolometers are the most widely used because of their advantages such as easy fabrication, stability, light-weight, ruggedness, and an easy array capability. To illustrate some principles of thermal detectors, consider the carbon bolometer shown in Fig. 8.9. Such devices are very sensitive and detect radiation over a very wide spectral range.

The resistance of an ordinary carbon resistor is a strong function of temperature, which makes a carbon resistor an inexpensive temperature sensor. To make a bolometer, we would mount the resistor in such a way that it is cooled to a low temperature but also isolated. As radiation strikes it, it warms up and the resistance decreases. An external electrical circuit detects the resistance change. To make the bolometer more sensitive, we would want to make its heat capacity smaller so that a small amount of energy could heat it faster.

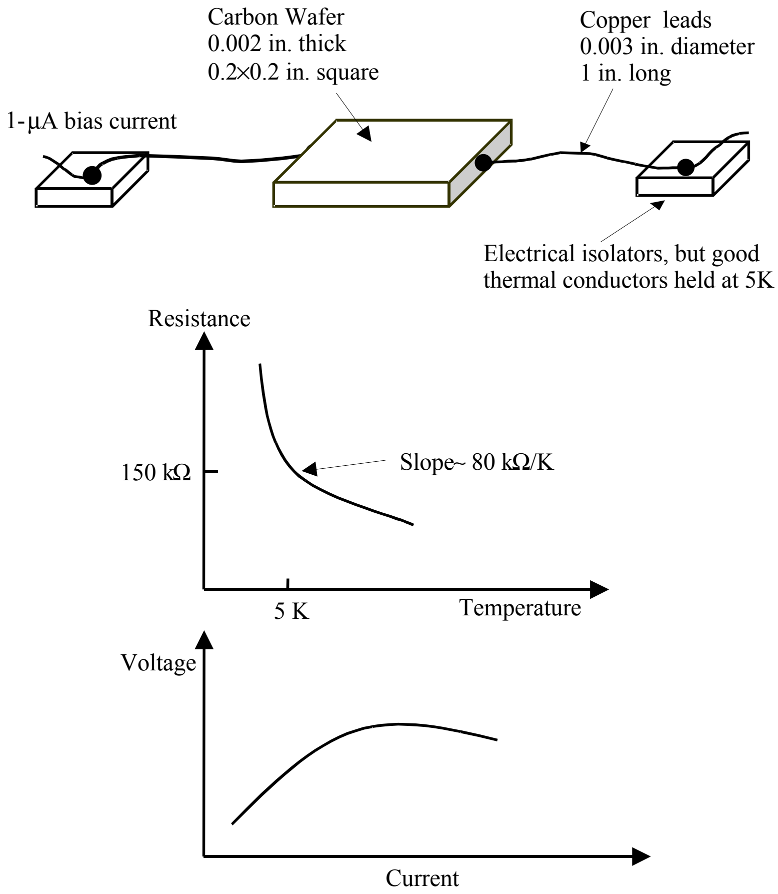

Carbon Wafer
0.002 in. thick
0.2×0.2 in. square
Copper leads
0.003 in. diameter
1 in. long
1-µA bias current
Electrical isolators, but good thermal conductors held at 5K
Resistance
150 kΩ
Slope~ 80 kΩ/K
5 K
Temperature
Voltage
Current

Fig. 8.9. Bolometer and its operating principle. A bolometer is a thermal detector whose resistance depends on its temperature. By precisely measuring the change in the resistance, one can determine how much radiant energy has reached the device.

When a modulated optical signal with power amplitude P_{in} and angular frequency ω hits a pixel with a heat capacity C, a temperature change ΔT can be recorded. The heat dissipation inside a solid is characterized by its thermal conductance K. If we define the quantum efficiency η as the fraction of the incident optical power that is absorbed by the solid, the temperature change is given by:

Eq. (8.26) $$\Delta T = \frac{\eta P_{in}}{K\left(1+\omega^2\tau_{th}^2\right)^{1/2}}$$

where τ_{th} is the thermal response time defined as the heat capacity of the material divided by its thermal conductance to the external equilibrium temperature:

Eq. (8.27) $$\tau_{th} = \frac{C}{K}$$

A few key points can be understood from Eq. (8.26) which can optimize the thermal detector performance. First, in order to sense low power infrared radiation it is very important to minimize the thermal conductance to the external environment K, by maximizing thermal isolation. Improved temperature isolation is often accomplished by lengthening or thinning the support legs in bolometer bridge structures, or by using thin wires which double as electrical contacts and suspension structure as shown in Fig. 8.9.". However, minimizing K also leads to a low frequency response, which is proportional to the heat dissipation rate. Secondly, the quality of the detector material is important in maximizing η, which is the ratio of incident to absorbed radiation. And lastly, the heat capacity C, of the element must be low enough in order to meet the response time requirement.

Recent research on bolometric arrays has demonstrated good room temperature performance. Wang et al. [2005] have deposited vanadium oxide on silicon substrates using ion beam sputtering. After micromachining, a bolometer array of 128 elements showed a detectivity of 2×10^8 cm.Hz^{½}.W^{-1} with a responsivity of 5 kV/W in the 8~12 μm regime. An SEM image of an array pixel is shown in Fig. 8.10.

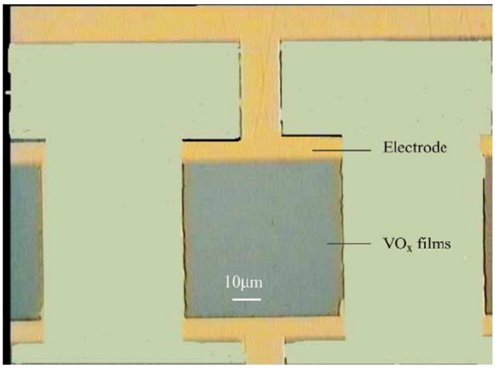


Fig. 8.10. Example of a bolometer array pixel. The active material is vanadium oxide. [Reprinted from Sensors and Actuators A Vol. 117, Wang, S.B., Xiong, B.F., Huang, G., Chen, S.H., and Yi, X.J., "Preparation of 128 element of IR detector array based on vanadium oxide thin films obtained by ion beam sputtering," p. 113, Copyright 2005, with permission from Elsevier.]

8.5. Summary

In this Chapter, the photodetector was established as a device that converts photon energy into electrical energy. The electromagnetic spectrum and blackbody concepts were initially revisited to provide a good background for the reader. The parameters used to describe the performance and limits of photodetectors were outlined, including responsivity, noise and detectivity. Finally, one specific type of photodetector, the thermal detector, and a carbon bolometer example were presented.

References

Johnson, J.B., "Thermal agitation of electricity in conductors," *Physical Review* 32, pp. 97-109,1928.

Jones, R.C., *Advances in Electronics*, Vol. V, Academic Press, New York, 1953.

Long, D., "On generation-recombination noise in infrared detector materials," *Infrared Physics* 7, pp. 169-170, 1967.

Rosencher, E. and Vinter B., *Optoelectronics*, Cambridge University Press, Cambridge, 2002.

Wang, S.B., Xiong, B.F., Zhou, S.B., Huang, G., Chen, S.H., Yi, X.J., "Preparation of 128 element of IR detector array based on vanadium oxide thin films obtained by ion beam sputtering," *Sensors and Actuators A* 117, pp. 110-114, 2005.

Further reading

Boyd, R.W., *Radiometry and the Detection of Optical Radiation*, John Wiley & Sons, New York, 1983.

Hudson Jr., R.D., *Infrared System Engineering*, John Wiley & Sons, New York, 1969.

Kingston, R.H., *Detection of Optical and Infrared Radiation*, Springer-Verlag, Berlin, 1978.

Problems

1. Sort the following types of electromagnetic radiation by order of increasing wavelength: radio, infrared, near-infrared, red, gamma rays, blue, ultraviolet, x-rays.

2. What is the total radiative emission of a surface (like a classroom wall) with area of 10 m^2, emissivity of 0.5 and temperature of 300 K? The

emissivity is simply a non-ideality factor for an object that does not emit and absorb perfectly. If this emission power seems high to you, remember that this surface is usually in thermal equilibrium and is also absorbing approximately the same amount of power from its surroundings.

3. Using Wien's law, calculate the peak wavelength of emission from these sources: The Sun (6000 K), tungsten filament (3000 K), red hot source (1000 K), 300 K ambient.

4. Name the following detector parameters:
   (a) Electrical output for a given light input.
   (b) The signal-to-noise ratio that would result if the performance of your detector were scaled to a detector of a standard size, under standardized test conditions.
   (c) The "clutter" or unwanted electrical variation that tends to hide the true signal.
   (d) The minimum infrared power that a detector can accurately "see".
   (e) A measure of the "cleanliness" of a signal pattern.
   (f) The condition when a detector's performance is not limited by intrinsic device noise, but rather the incident photon noise.

5. Assume you have a photodetector with a current responsivity of 5 A/W. A continuous wave HeNe laser with wavelength of 632.8 nm and spot size of 0.5 mm^2 is incident on the active area of the detector. Assume that 6.37×10^{22} photons per m^2 arrive on the detector each second. Calculate the expected photocurrent under these conditions.

6. A coaxial cable is used to transmit data with a bandwidth of 100 MHz. Calculate the maximum peak Johnson noise power for 300 K and 80 K.

7. A certain HgCdTe detector has a specific detectivity of $1 \times 10^{11}\ \text{cm}\cdot\text{Hz}^{1/2}\cdot\text{W}^{-1}$ and is designed to operate at 10 μm. It is incorporated into a focal plane array that needs to operate with a bandwidth of 1 kHz. This particular detector has a pixel size of 30 μm × 30 μm. Calculate the minimum power that can be sensed by this detector pixel.

8. You are tasked with designing a night vision infrared imaging system that should be able to detect a incident signal power level of $1 \times 10^{-7}\ \text{W}\cdot\text{cm}^{-2}$. Assume that you need to use a focal plane array with pixels of 30 μm × 30 μm operating at 60 Hz. What minimum specific detectivity does your system need to have?

9. Describe briefly how thermal detectors work.

10. What specific property changes with temperature in the following thermal detectors?
    (a) mercury-in-glass thermometer
    (b) carbon bolometer
    (c) thermopile
    (d) Golay cell

9. Photon Detectors

9.1. **Introduction**
9.2. **Types of photon detectors**
  9.2.1. Photoconductive detectors
  9.2.2. Photovoltaic detectors
9.3. **Examples of photon detectors**
  9.3.1. *p-i-n* photodiodes
  9.3.2. Avalanche photodiodes
  9.3.3. Schottky barrier photodiodes
  9.3.4. Metal-semiconductor-metal photodiodes
  9.3.5. Type II superlattice photodetectors
  9.3.6. Photoelectromagnetic detectors
  9.3.7. Quantum well intersubband photodetectors
  9.3.8. Quantum dot infrared photodetectors
9.4. **Focal Plane Arrays**
9.5. **Summary**

9.1. Introduction

In the previous Chapter, the basic concepts of photodetectors were outlined. Furthermore, thermal detectors and the bolometer, specifically, were described in detail. In photon detectors, incident photons interact with the electrons in the material and change the electronic charge distribution. This perturbation of the charge distribution generates a current or a voltage that can be measured by an electrical circuit. Because the photon-electron interaction is "instantaneous", the response speed of photon detectors is much higher than that of thermal detectors. Indeed, by contrast to thermal detectors, quantum or photon detectors respond to incident radiation through the excitation of electrons into a non-equilibrium state. The mechanisms of electron excitation are shown in Fig. 9.1.

Semiconductor photon detectors may rely on interband electron excitation (Fig. 9.1(a)) (intrinsic detectors), on impurity-band transition (Fig. 9.1(b)) (extrinsic detectors) or intersubband transitions in a quantum

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_9,
© Springer Science+Business Media, LLC 2010

well (Fig. 9.1(c)) (quantum well intersubband photodetectors). In intrinsic semiconductors, an electron in the valence band absorbs the energy of an incident photon and is excited into the conduction band. Extrinsic semiconductor detection occurs when an electron from a trap level inside the bandgap absorbs the energy of an incident photon and is excited into the conduction band. Intersubband detectors feature an electron in a quantum-confined state that can be excited into a higher energy state or continuum level. In the following sub-sections, the two main types of photodetectors, photoconductive and photovoltaic detectors, will be discussed in more detail.

The output signal of a photon detector due to incident light is called the photoresponse. It is strongly dependent on the frequency or wavelength of the incident light. When the wavelength of the incident light becomes longer than a critical wavelength, the photoresponse decreases abruptly. This particular wavelength generally corresponds to the bandgap of the semiconductor material in intrinsic detectors or to the activation energy of defect states in extrinsic detectors. For those wavelengths longer than this critical point, the energy of the incident photons is no longer sufficient to excite an electron-hole pair across the bandgap or to overcome the activation energy. The wavelength at which this abrupt decrease in responsivity occurs is called the cut-off wavelength.

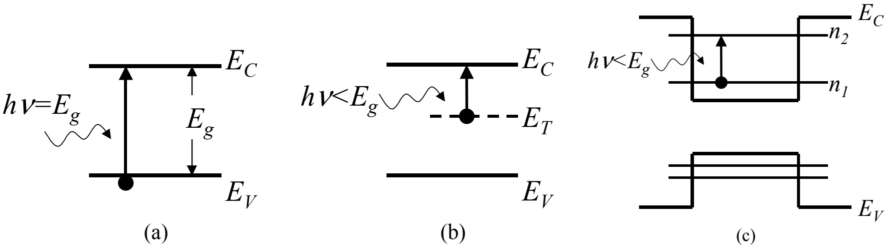


Fig. 9.1. Different mechanisms of excitation of an electron: (a) intrinsic semiconductor; (b) extrinsic semiconductor; and (c) intersubband transitions in a quantum well.

For example, InSb has a bandgap of 0.22 eV, which corresponds to a wavelength of 5.6 μm. Photons with longer wavelengths will pass through InSb undetected, i.e. InSb is transparent in spectral region beyond 5.6 μm, while photons of wavelengths shorter than 5.6 μm are effectively absorbed by InSb and contribute to the responsivity. Thus we expect to see a cut-off wavelength of 5.6 μm. This property is true for both photoconductors and photovoltaic detectors.

Examples of photoconductive detectors include doped germanium (Ge:X) and silicon (Si:X), and lead salts (PbS, PbSe). The ternary compounds HgCdTe and PbSnTe can be used as photoconductors and also as photovoltaic detectors.

9.2. Types of photon detectors

9.2.1. Photoconductive detectors

A photoconductive detector (also called a photoconductor) is essentially a radiation-sensitive resistor. The operation of a photoconductor is shown in Fig. 9.2. A photon of energy $h\upsilon$ greater than the bandgap energy E_g is absorbed to produce an electron-hole pair, thereby changing the electrical conductivity of the semiconductor. In almost all cases, electrodes attached to the sample measure the change in conductivity. Photoconductors are usually biased using a battery and a load resistor. An increase in the detector conductance both increases the current and decreases the voltage across the detector. For low resistance material, the photoconductor is usually operated in a constant current circuit as shown in Fig. 9.2. The series load resistance is large compared to the sample resistance, and the signal is detected as a change in the voltage developed across the sample. For high resistance photoconductors, a constant voltage circuit is preferred, and the signal is detected as a change in current.

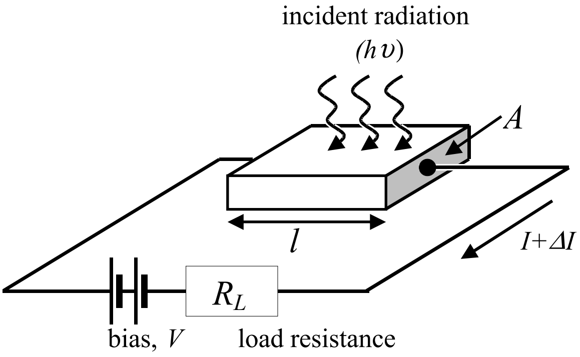


Fig. 9.2. Photoconductor and its biasing circuit.

The photoconductivity $\Delta\sigma$ is the difference between the electrical conductivity when a photoconductive material is illuminated and non-illuminated:

Eq. (9.1) $$\Delta\sigma = q(\mu_e \Delta n + \mu_h \Delta p)$$

where Δn and Δp are the excess electron and hole concentrations resulting from the incident light, μ_n and μ_p are the electron and hole mobility, respectively. In this case, the photoresponse takes the form of a change in the electrical current flowing through the sample, as a result of the change in electrical conductivity. The additional current component is called photocurrent. In a sample with cross-sectional area A and length l, the photocurrent under the bias V is given by:

Eq. (9.2) $$\Delta I = q(\mu_e \Delta n + \mu_h \Delta p)\frac{A}{l}V$$

The excess carrier concentrations generated under a steady-state illumination are:

Eq. (9.3) $$\Delta n = \Delta p = G\tau_n$$

where τ_n is the recombination lifetime of the excess carriers, and G is the excess carrier generation rate. This quantity is further related to the incident optical power P_{in} through:

Eq. (9.4) $$G = \eta \frac{P_{in}}{h\upsilon}\frac{1}{lA}$$

where the quantum efficiency η represents the fraction of the incident optical power that contributes to electron-hole pair generation, and υ is the frequency of the incident light. Eq. (9.2), Eq. (9.3) and Eq. (9.4) can be combined and give:

Eq. (9.5) $$\Delta I = q\left(\frac{\eta P_{in}}{h\upsilon}\right)\tau_n(\mu_e + \mu_h)\frac{V}{l^2}$$

For $\mu_e >> \mu_h$, this expression becomes:

Eq. (9.6) $$\Delta I = q\left(\frac{\eta P_{in}}{h\upsilon}\right)\frac{\tau_n}{\tau_t}$$

where we have defined the quantity:

Eq. (9.7) $$\tau_t = \frac{l^2}{\mu_e V}$$

which physically represents the electron transit time across the electrodes, i.e. the time taken by an electron to travel or transit from one end of the semiconductor to the other, separated by a distance l, when a bias V is applied. The ratio $\frac{\tau_n}{\tau_t}$ in Eq. (9.6) characterizes how fast the electrons can transit from one electrode to another electrode and contribute to the photocurrent before recombination occurs. It is called the photoconductive gain.

The current responsivity R_i of a photoconductor is defined as the ratio of the output signal, i.e. the photocurrent ΔI, to the input signal, i.e. the incident optical power P_{in}:

Eq. (9.8) $$R_i = \frac{\Delta I}{P_{in}} = q\left(\frac{\eta}{h\upsilon}\right)\frac{\tau_n}{\tau_t}$$

The units of the current responsivity are $\mathrm{A \cdot W^{-1}}$. It is also common to express the responsivity as a function of the wavelength λ of the incident light:

Eq. (9.9) $$R_i = \frac{\Delta I}{P_{in}} = q\left(\frac{\eta\lambda}{hc}\right)\frac{\tau_n}{\tau_t}$$

When the incident optical power is modulated by a sinusoidal signal of frequency ω, the photoresponse can be considered in terms of the root-mean-square (RMS) of the photocurrent, i.e.:

Eq. (9.10) $$I_{rms} = q\left(\frac{\eta P_{rms}}{h\upsilon}\right)\frac{\tau_n}{\tau_t}\frac{1}{\sqrt{1+(\omega\tau_n)^2}}$$

where $P_{rms} = \frac{P_{in}}{2}$ is the RMS of the incident optical power and P_{in} is the maximum amplitude of the sinusoidally modulated power.

Depending on the electrical circuit considered, the photoresponse can sometimes be expressed as the ratio of the output voltage to incident optical power. The choice of the output signals for a photoconductor, either current or voltage, generally depends on the application in which the photodetector will be used. As shown in Eq. (9.9), the responsivity is a linear function of the wavelength of the incident light, up until the cut-off wavelength is reached. Beyond that wavelength, the responsivity abruptly decreases as discussed previously.

9.2.2. Photovoltaic detectors

Photoconductors are passive devices that need an external electrical bias in order to operate and do not generate a voltage by themselves. They can consist of a simple block of semiconductor material. By contrast, a photovoltaic detector needs a more complex structure that uses a *p*-*n* junction. Such a detector is also called a photodiode. This allows the detector to exhibit a voltage when photons are absorbed as we will now briefly discuss. As a result, photovoltaic detectors are usually operated with a low or zero external bias.

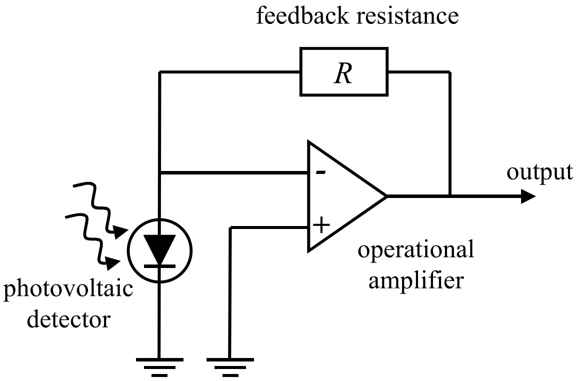


Fig. 9.3. Photovoltaic detector circuit using an operational amplifier in the feedback mode. When incident radiation is absorbed by the photovoltaic detector, a voltage is generated which is then collected through the circuit shown above.

In such a detector, as a result of the built-in electric field in the depletion region, the photogenerated carriers drift to opposite sides of the depletion region: holes toward the *p*-type side and electrons toward the *n*-type side. There, they increase the majority carrier densities on both sides of a junction. An open-circuit voltage generated by this build-up of charge can then be measured. Fig. 9.3 depicts an example of the electrical circuit commonly used with photovoltaic detectors. No specific bias circuit is necessarily used in the photovoltaic detector operation.

The simple calculation conducted in the case of a photoconductor can be easily applied in the case of a photovoltaic detector to obtain an expression for the current responsivity:

Eq. (9.11) $$R_i = \frac{q\lambda\eta}{hc}$$

where the parameters have the same meaning as defined earlier.

A more precise calculation can be conducted as follows. Since we are considering a *p-n* junction, we can make use of the diode equation, which relates the current to the applied external voltage as obtained in Eq. (4.56) in Chapter 4. The current-voltage characteristic of a photovoltaic detector is illustrated in Fig. 9.4. The current given by this equation is termed the dark current, i.e. the current which would be flowing through the device without illumination. To obtain the total current across the detector, we must add to the dark current the photocurrent which is the component directly resulting from the photogenerated electron-hole pairs. The total current is then given by:

Eq. (9.12) $$I = I_0\left(e^{\frac{qV}{k_bT}} - 1\right) - I_{ph}$$

where V is applied external voltage, I_0 is the saturation current, and I_{ph} is the photocurrent, which is also called the short-circuit current and has the following expression:

Eq. (9.13) $$I_{ph} = qAG(L_n + L_p)$$

where A is the cross-sectional area of the *p-n* junction diode. G is the excess carrier generation rate. The gain of the device is usually considered to be unity for photodiodes, L_n and L_p are the diffusion lengths of electrons and holes, respectively. The current-voltage characteristic of a photodiode under varying degrees of illumination is depicted in Fig. 9.4.

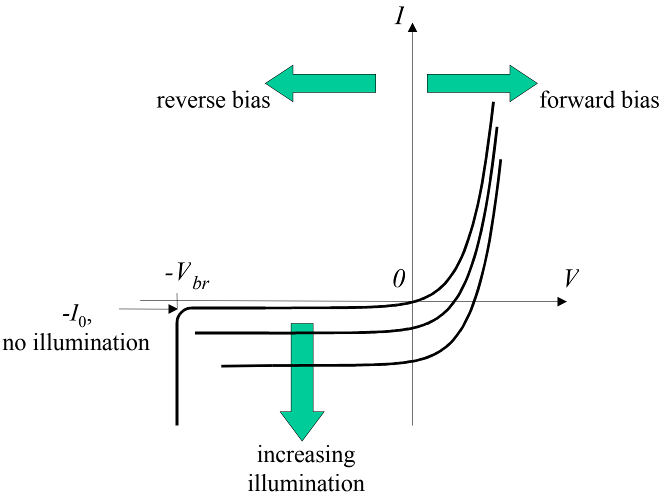


Fig. 9.4. Current-voltage characteristic of a photovoltaic detector with and without illumination.

The voltage that could be measured across the photodetector in an open-circuit situation can be found from Eq. (9.12) by setting $I = 0$:

Eq. (9.14) $$V_{oc} = \frac{k_b T}{q} \ln\left(1 + \frac{I_{ph}}{I_0}\right)$$

Since a photovoltaic detector can operate without external voltage, one important characteristic of the detector is its differential resistance R_0 at zero bias, defined by:

Eq. (9.15) $$\frac{1}{R_0} = \left(\frac{dI}{dV}\right)_{V=0}$$

By differentiating Eq. (9.12) with respect to the voltage V, and calculating it at $V = 0$, we can express the saturation current as a function of the differential resistance R_0:

Eq. (9.16) $$\frac{1}{R_0} = \frac{qI_0}{k_b T}$$

Using Eq. (9.13), we can express the ratio of the photocurrent to the saturation current which appears in Eq. (9.14):

Eq. (9.17) $$\frac{I_{ph}}{I_0}=\frac{qAG(L_n+L_p)}{k_bT/qR_0}=\frac{q^2G(L_n+L_p)}{k_bT}R_0A$$

We therefore observe that this ratio is proportional to the product R_0A, which is a useful figure of merit for photovoltaic detectors.

A more simple analytical expression for the detectivity than the one given in Eq. (8.21) can be obtained in the case of a *p-n* junction photovoltaic detector when the thermal noise is dominant over all other noise sources:

Eq. (9.18) $$D^*=\frac{q\eta}{h\upsilon}\sqrt{\frac{R_0A}{4k_bT}}$$

in which all the terms have been defined previously. This equation directly relates the R_0A product to the thermally-limited detectivity.

Si, InSb, and HgCdTe are examples of materials commonly used for photovoltaic infrared photodetectors. Some of the advantages of photovoltaic detectors over photoconductive ones include a better theoretical signal-to-noise ratio, simpler biasing, and a more predictable responsivity. However, photovoltaic detectors are generally more fragile than photoconductors. Indeed, they are susceptible to electrostatic discharge and to physical damage during handling. In addition, because they are thin (10 µm for a backside illuminated InSb *p-n* junction), the insulating layers are susceptible to electrical breakdown. Surface effects may also lead to leakage between the *p*-type and the *n*-type regions, which can then degrade detector performance.

9.3. Examples of photon detectors

In addition to the simple photoconductive and *p-n* junction photovoltaic detectors discussed previously, there are other types of photon detectors which will be briefly described in this section, including the *p-i-n* photodiode, avalanche photodiode (APD), Schottky photodiode, photoelectromagnetic (PEM) detectors, and quantum well and dot detectors.

9.3.1. p-i-n photodiodes

A *p-i-n* photodiode consists of a *p-n* junction diode within which an undoped intrinsic or *i*-region is inserted between the doped regions. Because of the very low density of free carriers in the *i*-region and its high resistivity, any applied bias is dropped almost entirely across this *i*-layer, which is then fully depleted at low reverse bias. The *p-i-n* diode has a "controlled" depletion layer width, which can be tailored to meet the requirements of photoresponse and frequency bandwidth. The absorption and carrier generation processes in a *p-i-n* photodiode are shown in Fig. 9.5.

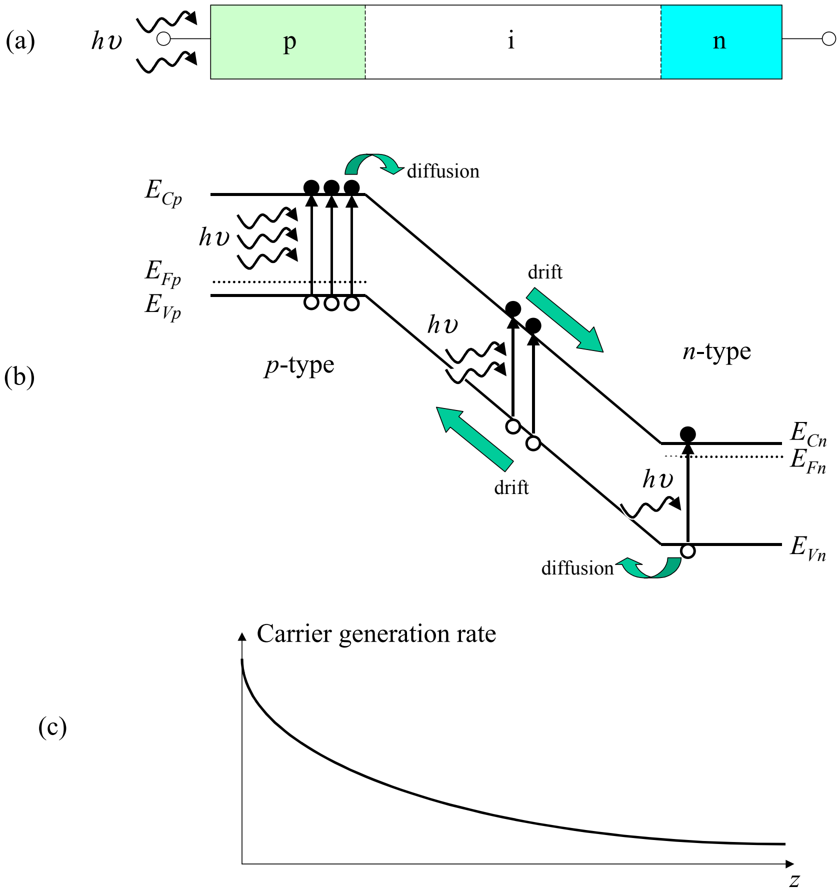

(a)
$h\upsilon$
p
i
n
(b)
E_{Cp}
diffusion
$h\upsilon$
E_{Fp}
E_{Vp}
p-type
$h\upsilon$
drift
n-type
drift
E_{Cn}
E_{Fn}
$h\upsilon$
E_{Vn}
diffusion
(c)
Carrier generation rate
z

Fig. 9.5. (a) Schematic structure of a reverse-biased p-i-n diode, with the incident light arriving on the p-type side; (b) the absorption of photons creates electron-hole pairs in the p-type, n-type and the i-region where they then become spatially separated through the electric field across the space charge region; and (c) graph of the carrier generation.

For practical applications, photoexcitation is provided either through an etched opening in the top contact, or an etched hole in the substrate, as schematically shown in Fig. 9.6. The latter reduces the active area of the diode to the size of the incident light beam.

By careful choice of the material parameters and device design, large bandwidths can be attained using *p-i-n* photodetectors. The response speed and bandwidth are ultimately limited either by the transit time or by circuit parameters. The transit time of carriers across the depletion or *i*-layer depends on its width and the carrier velocity and can be reduced by making the *i*-layer more thin, at the possible expense of reducing the overall photosignal.

The key elements in achieving high-performance with a *p-i-n* diode (high quantum efficiency and large bandwidth) is to illuminate the diode through the substrate, also called back-side illumination, ensure the total depletion of the *i*-layer, and use the device at a low reverse bias. The latter is important for digital operation and for low-noise performance.

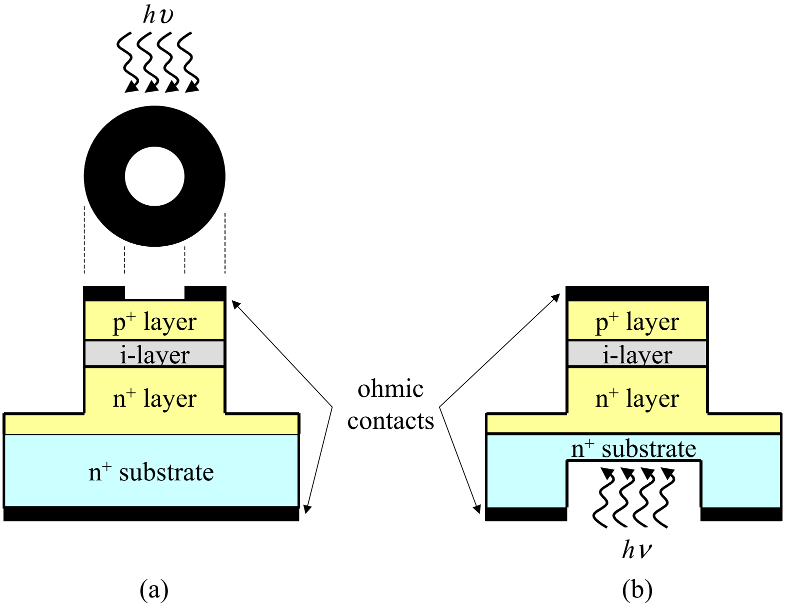


Fig. 9.6. Examples of mesa-etched p-i-n photodiodes for (a) top illumination and (b) back illumination. Top and back illuminations refer to the direction of the incident radiation to be detected with respect to the substrate on which the photodiode is fabricated.

9.3.2. Avalanche photodiodes

An avalanche photodiode (APD) operates by converting each absorbed photon into a cascade of electron-hole pairs. The device is a strongly reverse-biased photodiode in which the junction electric field is large. The charge carriers therefore accelerate in the space charge region, acquiring enough energy to generate additional electron-hole pair through impact ionization. This phenomenon is discussed in greater detail in Chapter 12.

The avalanche multiplication process is illustrated in Fig. 9.7. A photon absorbed at point 1 creates one electron-hole pair. The electron accelerates in the strong electric field. The acceleration process is constantly interrupted by random collisions with the lattice in which the electron loses some of its acquired energy. These competing processes cause the electron to reach an average saturation velocity. When the electron can gain enough kinetic energy to ionize an atom, it creates a second electron-hole pair. This is called impact ionization (at point 2). The newly created electron and hole both acquire kinetic energy from the electric field and create additional electron-hole pairs (e.g. at point 3). These in turn fuel the process, creating other electron-hole pairs. This process is therefore called avalanche multiplication.

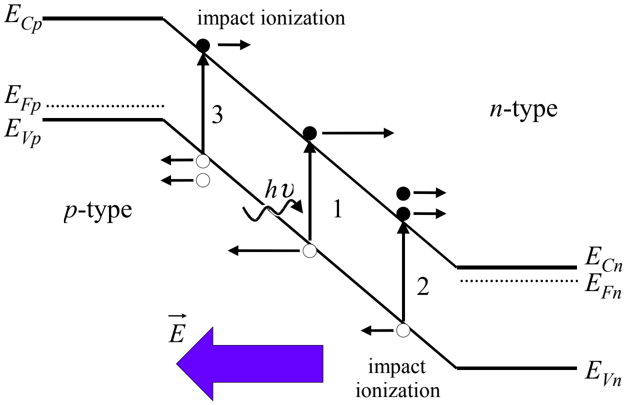


Fig. 9.7. Schematic representation of the multiplication process in an avalanche photodiode.

The abilities of electrons and holes to ionize atoms are characterized by their ionization coefficients α_e and α_h. These represent the ionization probabilities for electrons and holes per unit length. The ionization coefficients increase with the electric field in the depletion layer and decrease as the device temperature is raised. An important parameter which characterizes the performance of an APD is the ionization ratio $k = \alpha_h/\alpha_e$. If holes do not ionize effectively ($k << 1$), most of the ionization events are caused by electrons. The avalanching process then proceeds principally from left to right, i.e. from the *p*-type side to *n*-type side, in Fig. 9.7. If electrons and holes both ionize appreciably ($k \approx 1$), the gain of the device, i.e. the total charge generated in the circuit per photogenerated carrier pair, increases. However, this situation is undesirable for several reasons: it is time consuming and therefore reduces the device bandwidth, it is a random process and therefore increases the device noise, and it may cause avalanche breakdown. It is therefore generally desirable to fabricate APDs from materials that permit only one type of carrier (either electrons or holes) to ionize effectively. The ideal case of single-carrier multiplication is achieved when k is 0 or ∞.

In an optimally designed photodiode, the geometry of the APD should maximize photon absorption and the multiplication region should be thin to minimize the possibility of localized uncontrolled avalanches.

Due to their speed and extreme sensitivity (they can easily count individual photons), avalanche photodiodes have been used extensively for commercial and military applications. Recently, APD devices have been used heavily in laser range finding or Light Detection And Ranging (LIDAR), which is a technique that allows measurements of distance, speed, rotation and even chemical composition and concentration. An aerial image from the LIDAR intensity data is shown in Fig. 9.8.

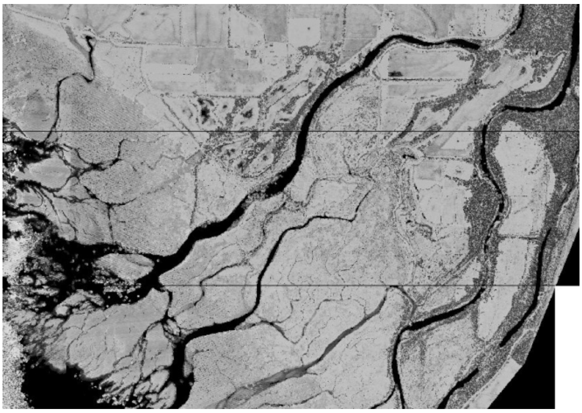


Fig. 9.8. LIDAR light intensity data. [From http://www.sbgmaps.com/lidar.htm. Reprinted with permission from Spencer B. Gross, Inc.]

9.3.3. Schottky barrier photodiodes

Schottky barrier photodiodes have been studied extensively and have found various applications. These devices have some advantages over *p*-*n* junction photodiodes such as their simplicity in fabrication, absence of high-temperature diffusion processes, and their high speed of response. The rectifying property of the metal-semiconductor junction, which is called a Schottky contact, has been reviewed in detail in sub-section 4.4.2. Briefly, the rectification arises from the presence of an electrostatic barrier between the metal and the semiconductor, which is due to the difference in work functions Φ_m and Φ_s of the metal and semiconductor, respectively, as shown in Fig. 9.9 for an *n*-type and a *p*-type semiconductor.

As also discussed in Chapter 4, the current transport across metal-semiconductor junctions is mainly due to the majority carriers, in contrast to *p-n* junctions where current transport is mainly due to minority carriers. It is now widely accepted that thermionic emission is the main process of carrier transport across Schottky barriers, and the current density is given by Eq. (A.4) in Appendix A.5. Knowing the current-voltage relationship, it is possible to calculate the R_0A product, as it was done in sub-section 9.2.2:

Eq. (9.19) $$\frac{1}{R_0A}=\frac{1}{A}\left(\frac{dI}{dV}\right)_{V=0}=\left(\frac{dJ}{dV}\right)_{V=0}$$

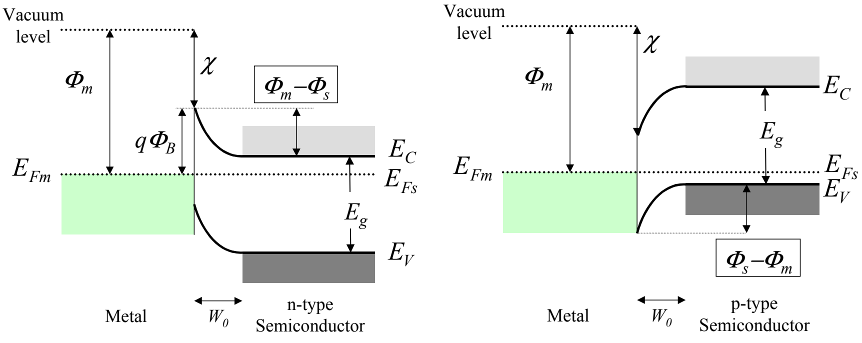
where A is the area of the Schottky junction. Using Eq. (A.4), we get:

Eq. (9.20) $$R_0A=\frac{k_bT}{qJ_{ST}}=\frac{k_b}{qA^*T}e^{\frac{q\Phi_B}{k_bT}}$$

where J_{ST} is the thermionic emission saturation current, k_b is the Boltzmann constant, A^* is the effective Richardson constant, and Φ_B is the Schottky potential barrier height for electron injection and is defined as:

Eq. (9.21) $$\Phi_B=\frac{\Phi_m-\chi}{q}$$

This quantity is illustrated in Fig. 9.9(a).


Vacuum level
Φ_m
χ
$\Phi_m-\Phi_s$
$q\Phi_B$
E_{Fm}
E_C
E_{Fs}
E_g
E_V
Metal
W_0
n-type Semiconductor
Vacuum level
Φ_m
χ
E_C
E_g
E_{Fm}
E_{Fs}
E_V
$\Phi_s-\Phi_m$
Metal
W_0
p-type Semiconductor

(a) (b)

Fig. 9.9. Equilibrium energy band diagram of Schottky contacts: (a) metal-(n-type) semiconductor ($\Phi_m>\Phi_s$); (b) metal-(p-type) semiconductor ($\Phi_m<\Phi_s$). A Schottky contact is obtained in each case because the majority carriers in the semiconductor experience a potential barrier which prevents their free movement across the metal-semiconductor junction.

9.3.4. Metal-semiconductor-metal photodiodes

A metal-semiconductor-metal (MSM) photodiode consists of two Schottky contacts on an undoped semiconductor layer. Unlike a *p-n* junction diode, it uses a planar structure. It can be designed such that the region between the metal fingers is almost completely depleted. When the semiconductor absorbs an incident photon, an electron-hole pair is created. The electron and the hole are spatially separated and accelerated under the influence of the applied electric field until they reach the metal contacts where they enter the biasing electrical circuit. An illustration of the energy band diagram for a MSM photodiode with an applied bias V is shown in Fig. 9.10.

The frequency response and bandwidth of a MSM photodiode are determined primarily by the transit time of the photogenerated carriers and the charge-up time of the diode. The MSM photodiode exhibits gain, has a low dark current, a large bandwidth, and is amenable to simple and planar integration schemes.

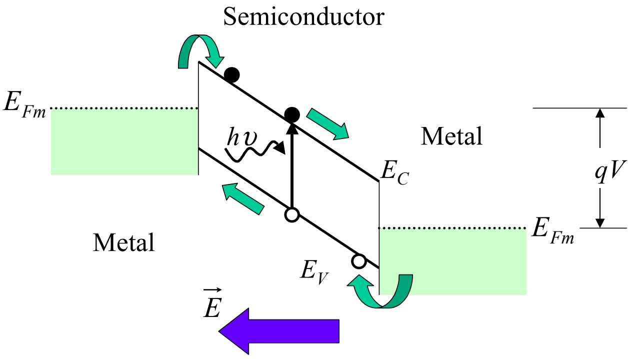


Fig. 9.10. Energy band diagram of an MSM photodiode under bias.

9.3.5. Type II superlattice photodetectors

The active layers of type II superlattice photodetectors are based on the type II band alignment in semiconductor heterojunctions where the conduction band of one layer lies below the valence band of the adjacent materials (see inset of Fig. 9.11). The basic physical properties behind such heterojunctions have already been discussed in sub-section 6.5.13 and have been illustrated in the case of semiconductor lasers.

A photoconductive InAs/GaSb type II superlattice detector and schematic diagram of the detection mechanisms inside the active region are shown in Fig. 9.11. Although the electrons and the holes are mostly confined in different layers, their wavefunctions can extend into the thin superlattice barriers. As a result, the overlap of the electron and hole wavefunctions is not strictly zero. The optical matrix element will have a high enough value to yield a considerable optical absorption only in the regions near the layer interfaces. Although this means a lower absorption than in type I quantum wells, the spatial separation of electrons and holes is advantageous in reducing the Auger recombination rate, which results in a longer lifetime of the photogenerated electron-hole pairs. This longer carrier lifetime, especially near room temperature, is the main advantage of this type of detector and has been experimentally demonstrated by Youngdale et al. [1994].

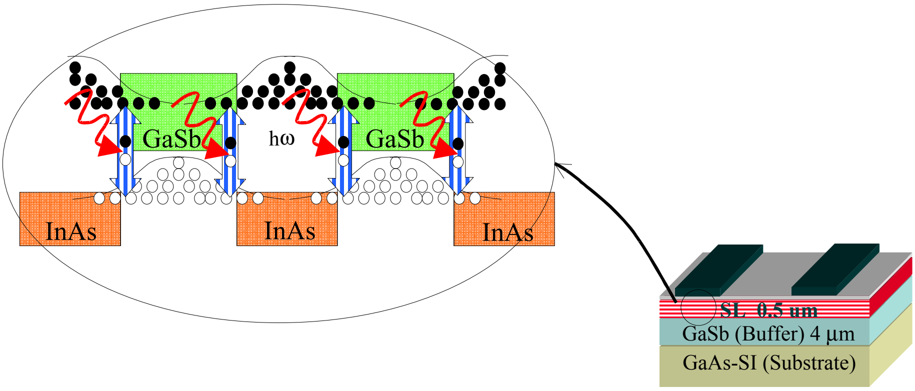


Fig. 9.11. Example of an InAs/GaSb type II photoconductive detector and the schematic diagram of the optical generation of electrons and holes in the active layer of the device.

Using type II superlattices as the narrow bandgap, active layer of a *p-i-n* photodiode, a detectivity of better than 10^{13} cm·Hz$^{1/2}$·W^{-1} has been recently achieved at 77 K with material designed for 5.4 µm cutoff wavelength [Walther et al. 2005]. Fig. 9.12 shows part of a focal plane array consisting of these devices. Other researchers at the Center for Quantum Devices have demonstrated good control of the cutoff wavelength of type II materials out to 32 µm [Wei et al. 2002]. At a wavelength of 8 µm, although the detectivity of HgCdTe is more than an order of magnitude higher than the type II photodetector, the latter still benefits from an easier growth process and a higher uniformity than HgCdTe, which can lead to less expensive focal plane arrays with comparable noise equivalent temperature difference (NEΔT) performance. The NEΔTis given by

Eq. (9.22) $$NE\Delta T = \frac{1}{(d/dT)\int_{\Delta\lambda} d\lambda (dR/d\lambda)} \frac{\sqrt{A\Delta f}}{D^*}$$

where the integral is over the spectral range of interest, the responsivity is R, the detectivity is given by D^*, the detector area is A, and Δf is the bandwidth.

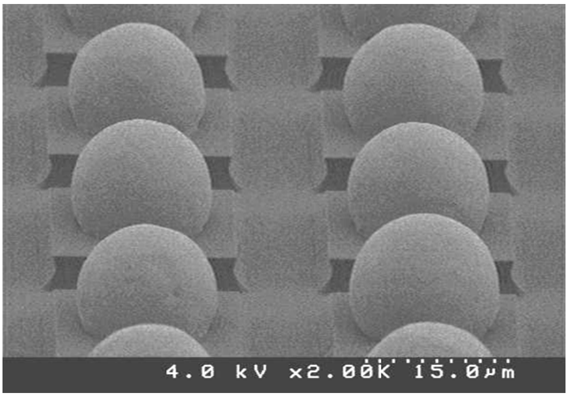


Fig. 9.12. Scanning electron microscope image of type II focal plane array pixels with indium bumps deposited on each pixel.

Currently, the only commercially available and fast room temperature photodetectors operating in the long wavelength infrared spectral region are based on HgCdTe and HgCdZnTe. In spite of their high detectivity, microbolometers have a time response which is at least three orders of magnitude slower than that of intrinsic detectors based on type II or HgCdTe. The new generation of HgCdZnTe detectors are now available with a special design to suppress the Auger recombination. Nevertheless, it has been shown that a type II superlattice with only 50 periods can have a higher detectivity at room temperature thanks to an lower Auger recombination rate (see Table 9.1 [Mohseni et al. 1998]).

Material	Type of detector	Wavelength (μm)	Detectivity ($\mathrm{cm \cdot Hz^{1/2} \cdot W^{-1}}$)
HgCdZnTe	Photovoltaic	10.6	1×10^7
HgCdTe	Photoelectromagnetic	10.6	5×10^6
HgCdZnTe	Photoconductive	10.6	6×10^6
Microbolometer	Thermal	8~14	5×10^8
Type II superlattice	Photoconductive	11	1×10^8

Table 9.1. Values of detectivity from typical infrared photon detectors at selected wavelengths at room temperature.

9.3.6. Photoelectromagnetic detectors

Photoelectromagnetic detectors operate on the principle of the photoelectromagnetic (PEM) effect. In a photoelectromagnetic detector a photon flux is incident upon a central window surrounded by a dark region that is hidden from illumination. The photon flux creates electron-hole pairs in the central illuminated region. Surface and bulk recombination effects in the dark area mean that these electron-hole pairs create a concentration gradient that drives diffusion of the locally photogenerated carriers into the surrounding dark region. A magnetic field is applied to the detector which causes a Lorentz force to act on the diffusion current. The deflection of these diffusing carriers in opposite directions due to a magnetic field, as shown in Fig. 9.13, causes a space charge build up, giving rise to an electric field directed along the x-axis (open-circuit voltage). If the sample ends are short circuited in the x-direction, current flows through the circuit (short-circuit current). The measured voltage or current can be directly related to the incident radiation which generated the carriers and the applied magnetic field:

Eq. (9.23) $$V_{PEM} = QB\frac{A_I}{d_D w}\Gamma$$

Where Q is the incident photon flux, B is the applied magnetic field, A_I is the illuminated area, d_D is the total width of the sensor, w is the thickness, and Γ is a constant dependant on the diffusion lengths and mobilities of carriers in the semiconductor.

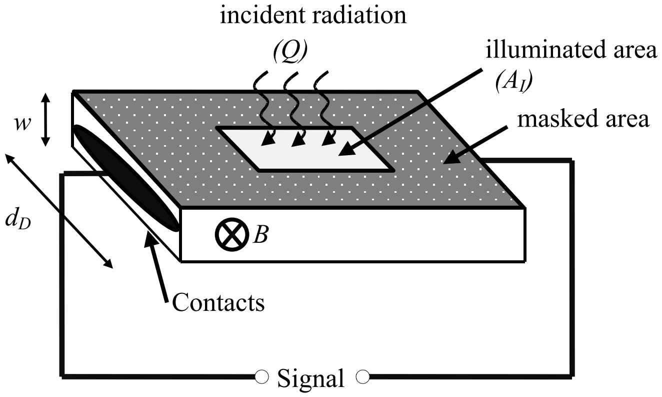


Fig. 9.13. Schematic of a detector based on the photoelectromagnetic effect.

Another way to achieve this situation is to use a graded gap system which makes the electrons and holes drift in the same direction or simply surface to bulk band gradient. The Lorentz force pushes them sideways in opposite directions and we have a net transverse electric field. Remember that unlike the Hall effect, here electrons and holes are being accelerated in the same direction by the graded gap, and thus deflect in the opposite direction. In the Hall case they move in opposite direction, deflect in the same direction, and the transverse Hall field would exactly cancel if they had the same mobility.

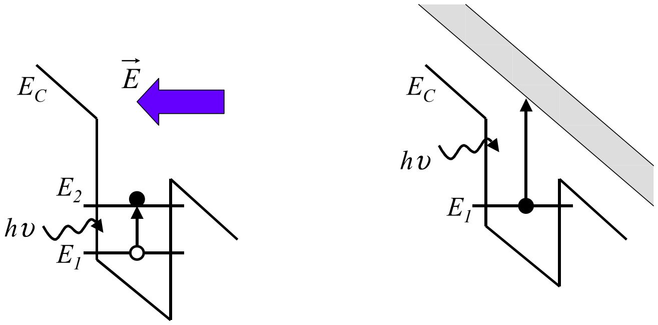
9.3.7. Quantum well intersubband photodetectors

A quantum well intersubband photodetector or (QWIP) is a device whose operation is based on the absorption of photons through the intersubband transition of carriers which are confined in multiple quantum wells.

QWIPs have a narrow absorption spectrum that can be tailored to match optical transitions in the 3~20 μm wavelength range by adjusting the quantum well width and barrier height or barrier layer composition. More importantly, it can be made using mature III-V semiconductors based on gallium arsenide (GaAs) or indium phosphide (InP) substrates.

The study of intersubband optical transitions in doped multiple quantum wells was motivated by the possibility of realizing high-speed infrared photodetectors. Both conduction band (*n*-type) and valence band (*p*-type) based quantum wells have been studied, although the larger hole effective mass results in poorer responsivity for the *p*-type devices.

The schematic for designing QWIPs is shown in Fig. 9.14(a). The absorption of photons having energy equal to the intersubband separation leads to transition of carriers between these subbands. For example, for III-V quantum wells of width $L = 100$ Å, the intersubband energy separation is in the range of 100~200 meV. The quantum mechanics selection rules allow absorption of electromagnetic radiation when the incident light polarization is parallel to the growth direction, i.e. TM polarization. This causes difficulties in detecting a two-dimensional image, since there is no effective absorption of light directed perpendicular to quantum well plane. However, illumination geometries using 45° facets as well as the use of surface gratings can circumvent the problem. For photoconductive devices utilizing intersubband absorption, the photogenerated carriers travel out of the quantum wells and contribute to the photocurrent.



(a) (b)

Fig. 9.14. Absorption of long-wavelength light in a quantum well due to (a) intersubband transitions in a wide well and (b) transition from a quasi-bound state to the continuum in a narrow well.

9.3.8. Quantum dot infrared photodetectors

Although quantum well intersubband photodetectors have been applied extensively in commercial and military applications, they suffer from low operating temperatures and non-normal incidence light absorption. Researchers have been developing quantum dot infrared photodetectors (QDIP) in order to overcome these disadvantages. In practice, several layers of III-V quantum dots are grown on lattice-mismatched matrix spacers and intersubband photon absorption takes place within the dots themselves. Due to the three-dimensional confinement of carriers, QDIP devices offer several potential advantages including the absorption of normally incident light. Furthermore, the quantum dots greatly reduce the rate of electron-phonon relaxation transitions, which leads to a longer average lifetime of photoexcited carriers and better overall device responsivity. Finally, lower dark currents may eventually allow high temperature operation.

One of the challenges of the QDIP architecture is in controlling the quantum dot size and shape. Most conventional fabrication techniques involve Stranski-Krastanov random growth in molecular beam epitaxy (MBE) or metalorganic chemical vapor deposition (MOCVD) and the dot geometry depends strongly on such parameters as growth temperature, V-III ratio, growth rate, etc. To add another set of variables, researchers can adjust the doping of quantum dot and/or surrounding epitaxial layers in order to control the carrier population of the device, for example.

Recent efforts in this detector architecture have resulted in room temperature operation with photoresponse out to 17 μm with a specific detectivity of 1.5×10^7 cm·Hz$^{1/2}$·W^{-1} and 1 V bias voltage at room temperature [Bhattacharya et al. 2005]. Future work in this field will probably involve improving the geometrical control of the nanometer-scale quantum dots as well as the periodicity and density of the dots.

9.4. Focal Plane Arrays

Solid state imaging systems consist of several different elements including the detector, optics, interconnects, and readout and clock circuitry. One type of detector architecture is the focal plane array (FPA). An FPA is an array of photodetectors placed at the focal plane of the imaging system's optics, which enables it to capture a two dimensional image. There are many application areas in which focal plane arrays are useful. These imagers find many uses in broadcast television, commercial photo and video devices (camcorders, for instance), machine vision, military and scientific applications. Arrays used for astronomy can boast pixel (individual image elements) numbers as large as a gigapixel, or one billion individual sensing elements!

Today, FPAs are available in monochromatic or multicolor systems, depending on the material type and wavelength range of interest. The most common types of imaging read-out architectures (essentially the manner in which the photosignal is handled within the device) include charge-coupled device (CCD) and complementary-metal-oxide-semiconductor (CMOS) arrays. One potential advantage to the CMOS design is the possibility of "per-pixel" signal processing, amplification and image correction.

Although focal plane array imagers are very common in our lives with products such as digital still and video cameras, they are quite complex to fabricate. Depending on the array architecture, the process can include over 150 individual fabrication steps. Contemporary visible imagers consist of silicon photodiodes integrated to an on-chip read-out integrated circuit (ROIC). When a detector substrate material different from the read-out circuit's silicon substrate is needed for different spectral regimes, the active sensing devices are often "hybridized" to the silicon-based read-out circuit. This hybridization process involves flip-chip indium bonding between the "top" surfaces of the ROIC and detector array. The indium bond must be uniform between each sensing pixel and its corresponding read-out element in order to insure high-quality imaging. After hybridization, a backside thinning process is usually performed to reduce the amount of substrate absorption. Some advanced FPA fabrication processes involve complete removal of the substrate material.

Fig. 9.15 shows two images taken using two different infrared FPAs. The left image was taken using a 256 × 256 hybridized infrared focal plane array with GaInAs/InP QWIP pixels operating at 8 μm [Jiang et al. 2003]. The right image was taken using a 256× 256 hybridized FPA with

InAs/GaSb type II superlattice pixels operating with a cutoff wavelength of 8 µm [Razeghi et al. 2005]. Both arrays were operated at liquid nitrogen temperature.

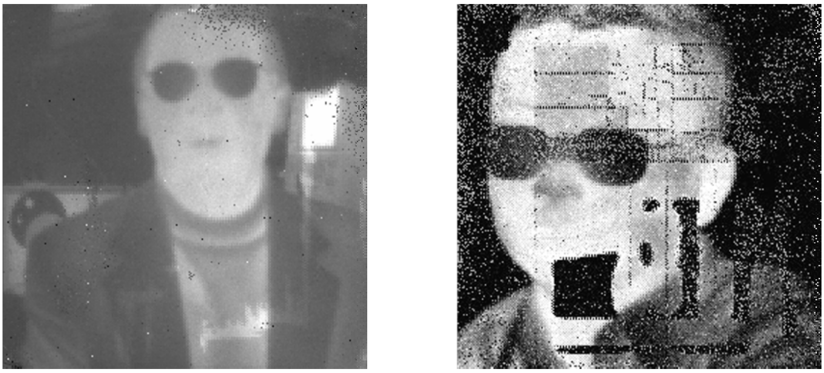

(a) (b)

Fig. 9.15. Images from a: (a) GaInAs/InP QWIP and (b) InAs/GaSb type II superlattice (right) focal plane array camera operating in the 8–12 µm wavelength regime.

9.5. Summary

In this Chapter, we have explored the topic of photon detectors. In particular, photoconductive and photovoltaic detector types were discussed in detail. Specific photodetector examples were also described, including *p-i-n*, avalanche, Schottky barrier, metal-semiconductor-metal, type II superlattice, and photoelectromagnetic detectors, as well as quantum well and quantum dot photodetectors. Finally, the topic of focal plane arrays was investigated, including a brief overview of the complex fabrication process.

References

Bhattacharya, P., Su, X.H., Chakrabarti, S., Ariyawansa, G., Perera, A.G.U., "Characteristics of a tunneling quantum-dot infrared photodetector operating at room temperature," *Applied Physics Letters* 86, pp. 191106-1, 2005.

Jiang, J., Mi, K., McClintock, R., Razeghi, M., Brown, G.J., and Jelen, C., "Demonstration of 256 × 256 focal plane array based on Al-free GaInAs-InP QWIP," *IEEE Photonics Technology Letters* 15, pp. 1273-1275, 2003.

Mohseni, H., Michel, E.J., Razeghi, M., Mitchel, W.C., and Brown, G.J., "Growth and characterization of InAs/GaSb type II superlattices for long wavelength infrared detectors," *Proceedings of the SPIE* 3287, pp. 30-37, 1998.

Razeghi, M., Wei, Y., Gin, A., Hood, A., Yazdanpanah, V., Tidrow, M.Z., and Nathan, V., "High performance Type II InAs/GaSb superlattices for mid, long, and very long wavelength infrared focal plane arrays," *Proceedings of the SPIE* 5783, pp. 86-97, 2005.

Wei, Y., Gin, A., Razeghi, M., Brown, G.J., "Advanced InAs/GaSb superlattice photovoltaic detectors for very long wavelength infrared applications," *Applied Physics Letters* 80, pp. 3262-3264, 2002.

Youngdale, E.R., Meyer, J.R., Hoffman, C.A., Bartoli, F.J., Grein, C.H., Young, P.M., Ehrenreich, H., Miles, R.H., and Chow, D.H., "Auger lifetime enhancement in InAs-$Ga_{1-x}In_xSb$ superlattices," *Applied Physics Letters* 64, pp. 3160-3162, 1994.

Further reading

Dereniak, E.L. and Crowe, D.G., *Optical Radiation Detectors*, John Wiley & Sons, New York, 1984.

Dereniak, E.L., and Boreman, G.D., *Infrared Detectors and Systems*, John Wiley & Sons, New York, 1996.

Holst, G.C., *CCD Arrays, Cameras and Displays, Second Edition*, SPIE Optical Engineering Press, Bellingham, WA, 1998.

Leigh, W.B., *Devices for Optoelectronics*, Marcel Dekker, New York, 1996.

Rogalski, A., *Infrared Photon Detectors*, SPIE Optical Engineering Press Bellingham, Washington, 1995.

Problems

1. What are the fundamental differences between a thermal and a photon (quantum) detector?

2. Assuming an interband absorption mechanism, what is the cutoff wavelength for a semiconductor with a bandgap of 0.2 eV? What is it if the bandgap is 0.02 eV?

3. Calculate the electron transit time and device gain for a photoconductor under these conditions: distance between contacts = 50 μm, applied voltage=10 V, minority carrier lifetime = 2×10^{-7} s, $\mu_n = 10000 \text{cm}^2/\text{V} \cdot \text{s}$, and $\mu_p = 1000 \text{cm}^2/\text{V} \cdot \text{s}$.

4. Describe briefly the spectral characteristics of the following detectors and explain the reasons for their particular spectral response shape.
   (a) InSb photovoltaic detector
   (b) Bolometer
   (c) Quantum well intersubband photodetector

5. Consider a hypothetical HgCdTe photodiode with an active area of 500 µm × 500 µm, $R_0 = 500\ \text{M}\Omega$, quantum efficiency = 0.6 operating at 80 K. Calculate the device specific detectivity at 8 µm wavelength.

6. List a few advantages and disadvantages of photovoltaic and photoconductive detectors.

7. Explain the possible advantages of a *p-i-n* photodiode over an abrupt *pn* design.

8. Describe the quantum selection rule as it pertains to *n*-type QWIP devices. How do conventional QWIP devices overcome this limitation? Can you list some more novel solutions to avoid this specific selection rule?

9. List the advantages and drawbacks of quantum well and quantum dot intersubband photodetectors.

10. Using knowledge gained from Chapter 3, describe the various processing steps that must be added to the conventional detector fabrication steps in order to generate a focal plane array. Assume pitch of 28 µm, InSb FPA with 640 × 320 pixels and Si ROIC. What are some of the complications added in the FPA hybridization process?

10. Type-II InAs/GaSb Superlattice Photon Detectors

10.1. Introduction
10.2. Material system and variants of Type II superlattices
  10.2.1. The 6.1 Angstrom family
  10.2.2. Type II InAs/GaSb superlattice
  10.2.3. Variants of Sb-based superlattices
10.3. Historic development of Type II superlattice photodetectors
10.4. Physics of Type II InAs/GaSb Superlattices
  10.4.1. Qualitative description
  10.4.2. Quantitative calculations of electronic bandstructure
10.5. Advantages of Type II superlattice
  10.5.1. Band gap engineering
  10.5.2. Auger suppression
  10.5.3. Large effective mass
  10.5.4. Normal incident, broad band absorption
  10.5.5. Good uniformity
10.6. Material growth and characterization
10.7. Device fabrication
  10.7.1. Single element device for testing
  10.7.2. Focal Plane Array Fabrication
10.8. Summary

10.1. Introduction

In the previous two chapters, the basic concept of photodetectors and examples for photon detector families were briefly described. Among the currently developing technologies, the only three that take advantages of low dimensional properties of quantum mechanics include: Type II InAs/GaSb superlattice photodetectors, the quantum well intersubband photodetectors and the quantum dot infrared photodetector. Going by dimensionality, the Type II superlattice is still a three dimensional system, the same as a bulk semiconductor, while the other two systems are so to

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_10,
© Springer Science+Business Media, LLC 2010

speak "two and zero dimensional systems" respectively, meaning that the confinements of carriers are locally in one or more than one direction. However, compared with the quantum well and the quantum dot, Type II superlattices share the same capability of spatially localizing electrons in a nanoscale space. This is due to the spatial separation of carriers in the constituent layers. Moreover, the physics that underlie the Type II superlattice give us the unique ability to control the interaction between localized electrons by way of varying potential barrier thickness. This gives us an additional degree of freedom in this system for designing and creating new "materials". This is not unlike the way Nature created solid crystals from atoms. Having bulk-like properties, Type II superlattices can be exploited using the considerable understanding acquired for bulk semiconductors in the past. This knowledge can be used to optimally design device architectures for photodetecting applications. The close match between theoretical work and experimental developments has allowed Type II superlattice photodetectors to become one of the fastest growing technologies in infrared detection. During a decade of maturing, the material system has experienced significant progress that has lifted the detector performance to now have comparable levels to the state-of-the-art Mercury Cadmium Telluride (MCT) technology. Whether the superlattice Infrared photodetectors can outperform the bulk narrow gap MCT detectors or not, is one of the most important questions for the future of the infrared detection technology.

In this chapter, we will explore the marvelous physics of the Type II superlattice. Quantum mechanical knowledge will be applied to qualitatively discuss the bandgap engineering of the superlattice structure. Then, taking this understanding forward, we will have a closer look at the fabrication of infrared photodetectors, and see how growth processes and device processing techniques are realized.

10.2. Material system and variants of Type II superlattices

10.2.1. The 6.1 angstrom family

The 6.1 Angstrom family (Fig. 10.1) is an important group in the semiconductor technology. It consists of three members that are closely lattice matched to each other: InAs (a = 6.0584 Å), GaSb (a = 6.0959 Å) and AlSb (a = 6.1355 Å). Since GaSb substrates became available, and the growth of III-antimonide semiconductors was feasible, the 6.1 Å and its compounds have been providing enormous flexibility in designing heterostructures for optical and electronic applications. One great advantage of the family, is the small lattice mismatch of the family to GaSb substrates and similar growth windows for the three materials. This enables the growth of high quality materials with a low density of defects and dislocations.

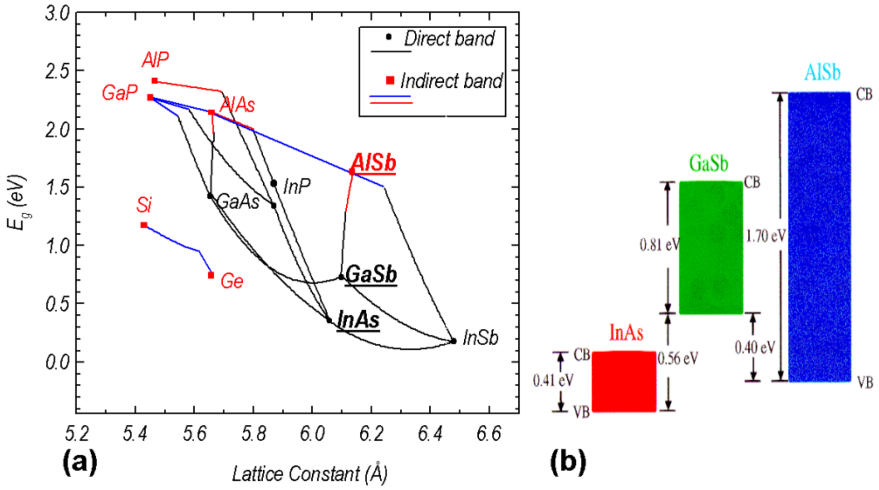


Fig. 10.1. (a) The energy gap versus the lattice constant of InAs, GaSb, and AlSb compared with other semiconductors. (b) InAs, GaSb, and AlSb energy band lineups.

The energy gap of the family and related compounds ranging from 0.41 eV (for InAs) to 1.70 eV (for AlSb) is of particular interest for optoelectronic devices in the Short Wavelength Infrared (SWIR) and Mid-Wavelength Infrared (MWIR) regimes. Moreover, the heterojunctions between InAs and the other two members benefit of the unique features of the Type II band alignment. On the one hand, the InAs/AlSb interface forms a Type II staggered line up where the conduction band of InAs is slightly above the valence band of AlSb. The high energy gap of AlSb leads to an exceptionally large conduction band off set of about 1.45 eV, enabling the realization of very deep quantum wells and very high tunneling barriers. This heterostructure has been widely utilized in high frequency field effect transistors (FETs) and resonant interband tunneling diodes (RITDs). On the other hand, the heterojunction between InAs and GaSb leads to the exotic so called *broken gap* line up, where the conduction band of InAs is about 0.15 eV lower than the valence band of Gasb. This type of *misaligned* structure is the reason why Type II superlattices have a flexible band gap engineering capability and this will now be discussed in the next sections.

10.2.2. Type II InAs/GaSb superlattice

The idea of Type II InAs/GaSb superlattice was first proposed by Sai-Halasz and Esaki in the 1970's. The superlattice is formed by alternating the InAs and GaSb layers over several periods. This creates a one dimensional periodic structure, like the periodic atomic chain in naturally occurring crystals (Fig. 10.2). The Type II broken gap alignment leads to the separation of electrons and holes into the InAs and GaSb layers, respectively. The charge transfer give rise to a high local electric field and strong interlayer tunneling of carriers without requiring an external bias or external doping. Large period superlattices behave like semimetals but if the superlattice period is shortened, the quantization effects are enhanced causing a transition from a semi-metal to a narrow gap semiconductor. The resulting energy gaps depend on the layer thicknesses and interface compositions. In reciprocal k-space, the superlattice is a direct bandgap material which enables optical coupling as shown in Fig. 10.2(b).

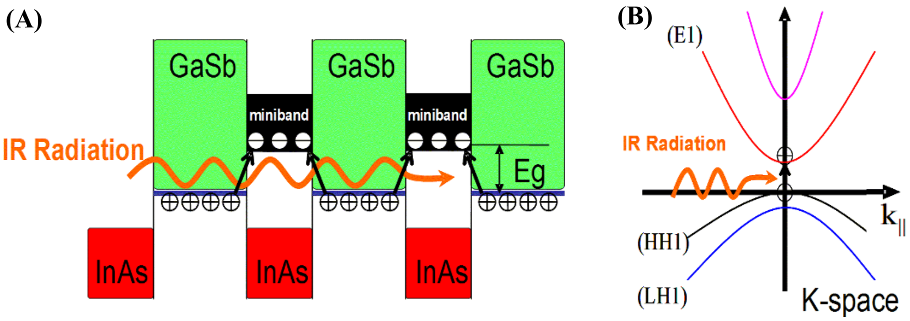


Fig. 10.2. a) Spatial band alignment in Type-II superlattice (red stands for InAs forbidden gap and green stands for GaSb forbidden gap) and b) band structure with direct band gap and absorption process in k-space

10.2.3. Variants of Sb-based superlattices

InAs/GaInSb

The first variant of the original Type II binary/binary InAs/GaSb superlattice is the binary/ternary InAs/GaInSb superlattice. The structure was proposed by Fuchs et al. [1997] to compensate for the spatial separation of electrons and holes, in order to enhance the optical matrix element. By adding indium to the GaSb material, the barrier height is reduced; the electron wave function becomes more delocalized, leading to a stronger overlap with the hole wavefunctions and a better transition probability. The GaInSb barrier also allows for one other degree of freedom in controlling the energy gap, i.e. via the change in the Ga/In molar fraction. This however also induces additional experimental difficulties because the growth of ternary GaInSb requires stricter control of growth temperature and III/V flux ratio than the growth of the binary GaSb. It is also more sensitive to the growth parameters, and results in a lower uniformity across a large growth area.

Another issue for ternary superlattices, is the positive net strain of the structure when it is grown on a GaSb substrate. As the lattice constant of InSb is much larger than that of GaSb, the GaInSb layer will make the superlattice structure more compressive (the lattice parameter of superlattice is larger than the lattice parameter of the substrate). Moreover, the lowering of the barrier due to the introduction of Indium into the GaSb layer tends to reduce the energy gap of the superlattice. When designing for a MWIR cut-off or shorter, the width of the InAs wells must be shortened in order to raise the electron energy level. This then increases the energy gap. The reduction of the InAs thickness will increase the average lattice constant of the superlattice, and induce even more strain that can further degrade the material quality.

W-structure

Contrary to the InAs/GaInSb superlattice, the W-structure superlattice (WSL) proposed by Meyer et al. is a design where the delocalization of electron is reduced by the insertion of GaAlSb layers. The W-structure is named after the characteristic W-shaped bandgap alignment in a single superlattice period, and was first used for quantum well lasers, and then applied for infrared photodiodes [Meyer et al. 1995]. The schematic diagram of the W-structure is presented in Fig. 10.3. The GaAlSb barrier was used to block the electron wavefunction overlap and increase the effective mass of the superlattice. In parallel, the thin GaInSb barrier is used to enhance the electron/hole interaction, as in the case of InAs/GaInSb superlattice. The disadvantages of the W-structure are again the difficulties associated with material growth. Similar to the ternary InAs/GaInSb superlattice, W-structure is based on ternary compounds and is sensitive to the growth condition. Moreover, the switching from the InAs layer to the GaAlSb layer means no common atom at the interface. This adds another degree of complexity in both theoretical calculations and experimental realizations. The large lattice mismatch between InAs and GaAlSb also limits the growth quality.

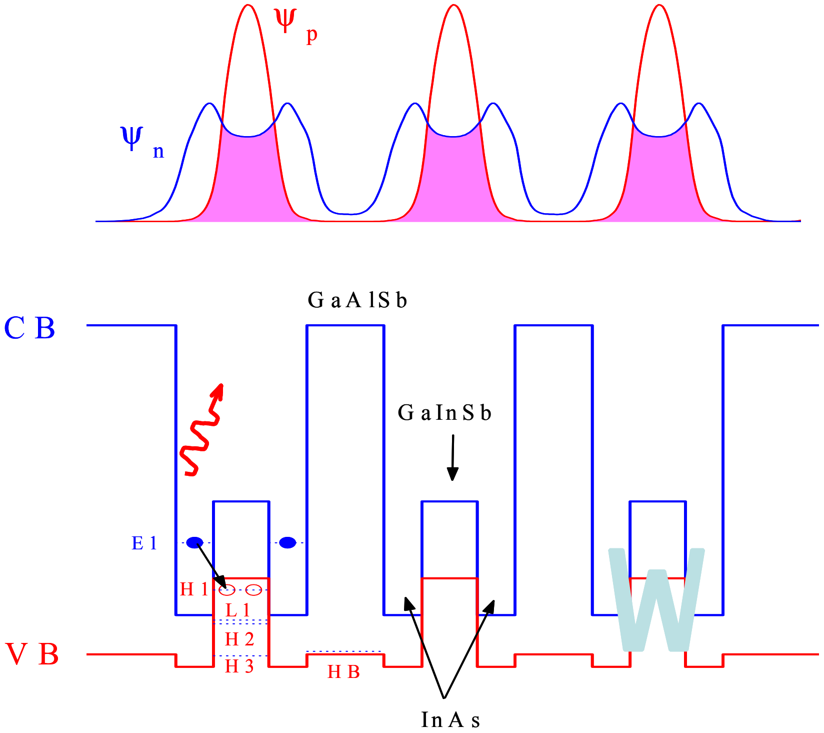


Fig. 10.3. Schematic diagram of the W-structure superlattice. a) Spatial Wavefunction of electron and hole; b) Spatial alignment of the lowest conduction band and the highest valence band of the constituent materials.

M-structure:

The M-structure superlattice, proposed by [Nguyen et al. 2007], is constructed by inserting a thin AlSb barrier in the middle of the GaSb layer of a normal Type II binary InAs/GaSb superlattice. Fig. 10.4 shows the schematic diagram of the energy band alignment of its constituents. The colored regions represent the prohibited band gap of the structure. This AlSb – containing superlattice is named the M-structure. This stands for the shape of the letter M of the band alignment of the AlSb/GaSb/InAs/GaSb/AlSb layers of a single superlattice period. The alignment allows the structure to acquire the beneficial properties of the well established binary – binary antimonide based SL structures used in high performance infrared detectors. First, it conserves the Type II misalignment of InAs and GaSb, which is capable of eliminating the electron-hole split-off band Auger transition by reducing the resonance between the energy gap and split-off band. Second, as illustrated in the Fig. 10.4, AlSb has a large energy bandgap with the conduction band offset higher than that of GaSb and the valence band offset slightly higher than that of InAs. Due to this band offset, AlSb can be utilized as a blocking layer for both electrons in the conduction band and holes in the valence band. As a consequence, the insertion of the AlSb into the GaSb, forming the M-superlattice, will reduce the tunneling probability in long wavelength (LWIR) and very long wavelength (VLWIR) photodiodes because the effective mass of the carriers over the standard binary-binary SL is increased. In comparison with other ternary/quaternary superlattices, such as the WSL or the InAs/GaInSb superlattices, the M structure is easier to realize because it still keeps the simple form of a binary-only structure, which is less sensitive to the growth temperature and the III-V flux ratio. Finally, the growth of the M-structure is as simple as a standard binary-binary InAs/GaSb structure because there is a small mismatch in the lattice constant between AlSb (6.136 Å) and GaSb (6.097 Å) and the common antimony atom allows for a smooth interface between the two layers. The M-structure does not need as much care as the growth of AlSb on InAs in the WSL where the interface of AlSb and InAs have no common atom and AlSb grown on InAs is highly strained.

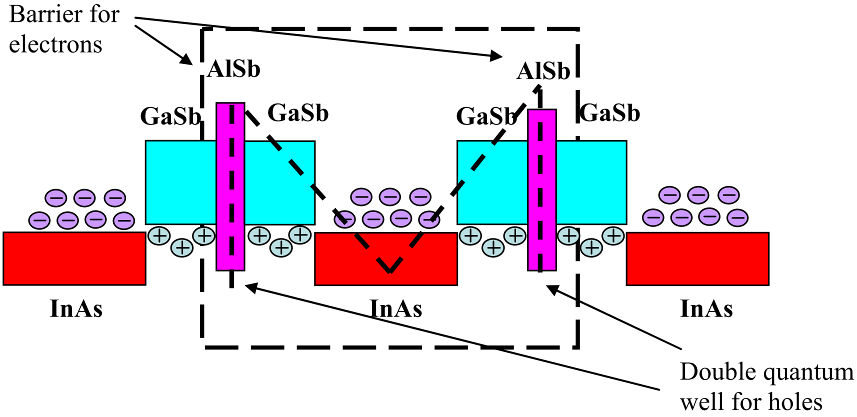


Fig. 10.4. Schematic diagram of the M structure. The inserted AlSb layer forms a barrier for electrons in the conduction band and a double quantum well for holes in the valence band. [Reprinted with permission from Proceedings of the SPIE Vol. 6479, Nguyen, B.M., Razeghi, M., Nathan, V., and Brown, G.J., "Type-II M structure photodiodes: an alternative material design for mid-wave to long wavelength infrared regimes," fig. 1, pg. 64790S-2, Copyright 2007, SPIE.]

10.3. Historic development of Type II superlattice photodetectors

In 1975, [Tsu et al. 1975] first proposed the concept of the semiconductor superlattice using an alternating GaAs/GaAlAs heterostructure. They identified the quantum states and predicted how this would give rise to the phenomenon of negative differential resistance (NDR) in the device. Two years later, the same group predicted important effects for superlattices with Type II alignment between two constituent layers in which, the conduction band of one layer lies below the valence band of the other [Sai-Halasz 1977]. Taking InAs/GaSb as a typical example for a type II superlattice, theoretical calculations were carried out to illustrate the energy gap shifting, and the semiconductor/semimetal transition when the layer thicknesses are changed. The structure was then proposed by [Smith et al. 1987] for infrared detection applications, exploiting the fact that the cut-off wavelength could be adjusted in the infrared range, and that the effective mass was constant, regardless of the energy gap.

Despite positive theoretical predictions, high-quality growth of this type of materials was not demonstrated until recent advances in the technology of the molecular-beam epitaxy (MBE) technique. The first experimental demonstration of photon detectors comprising superlattices was by [Johnson et al. 1996]. The first photoconductors grown on GaAs substrate were demonstrated at the Center for Quantum Devices (CQD) at Northwestern University in 1997 [Mohseni et al. 1997]. However, Type II superlattices grown on GaAs substrate experienced many difficulties due to the 7% lattice mismatch between the epitaxial layer and the substrate. Since epi-ready GaSb substrates became commercially available in the late 1990s, the growth of Type II superlattice on lattice matched GaSb have been significantly improved. Defect and dislocation free crystals enable the realization of photodiode detectors which require large differential resistance and low leakage. In the perpendicular *p-i-n* superlattice photodiode configuration, the signal current flows in the growth direction, the size of the detector element could thus be reduced, and multi-element devices, such as focal plane arrays (FPAs) became feasible. First high performance photodiodes using ternary InAs/GaInSb superlattices were demonstrated by F. Fuchs et al. in Germany in 1997 [Fuchs et al. 1997]. Since then, efforts have been concentrated in the optimization of photodiode performance and the fabrication of FPAs. In 2003, the first demonstration of an FPA based on Type II superlattice material systems was carried out at the CQD. This provided the proof of concept of this technology. One year later, a high performance FPA with 5 μm cutoff was demonstrated in 2004 in Germany [Cabanski et al. 2004].

Recently, excellent material quality has been achieved with matured growth techniques, allowing for the realization of complex device architectures with higher performance. [Vurgaftman et al. 2006] proposed graded gap, and then hybrid graded gap designs for photodiodes based on W-structure superlattice, and have proved that the dark current due to tunneling and recombination processes in LWIR photodiodes can be suppressed. [Nguyen et al. 2007] have applied the great tunability of the M-structure superlattice for the p^+-π-M-n^+ design that also exhibited an even greater suppression of the tunneling current in LWIR and VLWIR photodiodes. Rodriguez et al. utilized the nBn design for Type II superlattice and proved the concept for MWIR single element detectors and FPA [Rodriguez et al. 2007]. Heterostructure designs with high bandgap contact regions were also independently proposed by [Delaunay et al. 2007] and by [Vurgaftman et al. 2006], to avoid the bending of energy bands near the device's sidewall, resulting in a significant reduction of surface leakage even on unpassivated devices. With high quality material, [Nguyen et al. 2007] have demonstrated a quantum efficiency as high as 50% in topside illumination configuration. [Delaunay et al. 2008] then showed that the same device could achieve a quantum efficiency up to 75% at backside illumination configuration, without antireflective coating. This value is equivalent to that the state-of-the-art Mercury Cadmium Telluride photodetectors. [Nguyen et al. 2008] then combined both the optical and electrical optimization schemes into one superlattice structure that exhibited a specific detectivity $8.1\times10^{11}\,\mathrm{cm{\cdot}Hz^{1/2}{\cdot}W^{-1}}$. The internal noise of the detector was even lower than the noise due to the fluctuation of the radiation background, and the performance was limited by the background radiation, and it was said to be a "Back ground limited infrared photodetector" (BLIP).

Side by side with the development of device performance, device processing techniques have been gradually improved, making the fabrication of high quality single element and FPAs more routine. Researchers at Fraunhofer have commercialized MWIR FPAs based on Type II superlattice, and in 2007, they demonstrated for the first time a MWIR-MWIR two color Type II superlattice FPAs [Walther et al. 2007]. In the LWIR range, Delaunay et. al realized FPA using M-structure superlattice design, and showed a reduction of dark currents by one order of magnitude compared to FPA using the conventional Type II InAs/GaSb *p-i-n* photodiode design. The historical progress in the development of this material system is summarized in Fig. 10.5

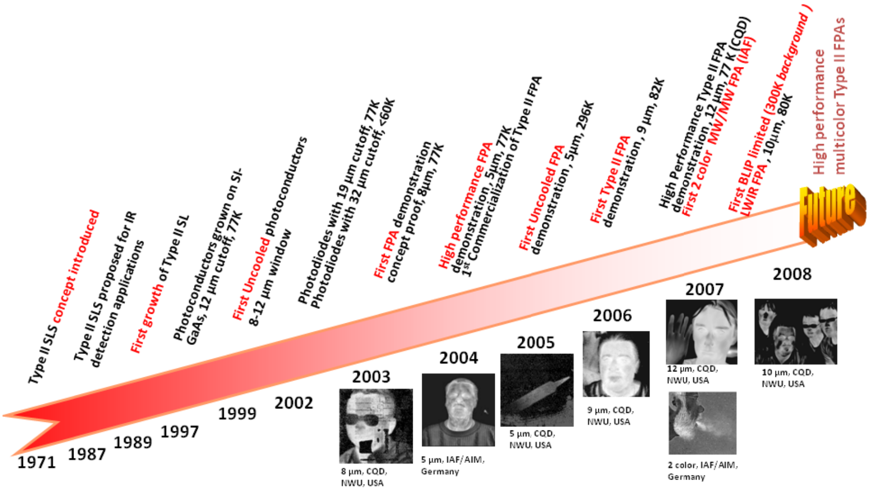

Type II SLS concept introduced
Type II SLS proposed for IR detection applications
First growth of Type II SL
Photoconductors grown on SI-GaAs, 12 μm cutoff, 77K
First Uncooled photoconductors 8-12 μm window
Photodiodes with 19 μm cutoff, 77K
Photodiodes with 32 μm cutoff, <60K
First FPA demonstration concept proof, 8μm, 77K
High performance FPA demonstration, 5μm, 77K
1st Commercialization of Type II FPA
First Uncooled FPA demonstration, 5μm, 296K
First Type II FPA demonstration, 9 μm, 82K
High Performance Type II FPA demonstration, 12 μm, 77 K (CQD)
First 2 color MW/MW FPA (IAF)
First BLIP limited (300K background) LWIR FPA, 10μm, 80K
High performance multicolor Type II FPAs
Future
1971
1987
1989
1997
1999
2002
2003
8 μm, CQD, NWU, USA
2004
5 μm, IAF/AIM, Germany
2005
5 μm, CQD, NWU, USA
2006
9 μm, CQD, NWU, USA
2007
12 μm, CQD, NWU, USA
2 color, IAF/AIM, Germany
2008
10 μm, CQD, NWU, USA

Fig. 10.5. Historical development of Type II superlattice photo-detectors.

10.4. Physics of Type II InAs/GaSb Superlattices

10.4.1. Qualitative description

In Type II superlattice, electrons and holes are spatially confined in the InAs and GaSb layers respectively. The electronic band structure of the superlattice is determined by the energy level of electrons and holes in the quantum wells, and by the interaction between the carriers in adjacent wells. In this section, we use quantum mechanics to qualitatively explain the dependence of the superlattice's energy gap on the layer thicknesses.

Effect of the InAs layer

The InAs layer is a quantum well confining the electrons in one period of the superlattice. As an empirical rule, the energy level of a particle is inversely proportional to its effective mass and the square of the well width. The conduction band level of Type II superlattice will lower when the InAs layer gets thicker. For example, a superlattice design aimed for a MWIR cut-off wavelength tends to have thin InAs layer (typically 4–10 Monolayer thick), while a LWIR and VLWIR structure normally has thicker InAs (typically 10–20 Monolayer thick). When the InAs layer is thicker, the energy level will be even lower than the valence band of the GaSb layer. In this case, electrons can directly tunnel from one well to another via the valence band of the GaSb barrier, the superlattice thus becomes a semimetal.

Effect of the GaSb layer

The GaSb layer acts as a well for holes, but also as the barrier isolating electrons in adjacent InAs wells. Since the effective mass of holes in the valence band is much heavier than that of electrons (about 1 order of magnitude larger than the electron effective mass), the dependence of the energy on the well width becomes much weaker. The energy level of holes with respect to the bottom of the well is normally very small. With a GaSb layer thicker than 1.5 nm (~5 Monolayers), the hole's energy level is practically unchanged, which means the valence band of superlattices stay constant near the valence band of the GaSb.

However, the GaSb thickness strongly affects the energy gap of the superlattice via the conduction band. Similar to the formation of bulk band structures from discrete atomic levels, the conduction band of the superlattice is the broadening of individual energy levels of electrons in InAs wells due to the "interaction" between wells. In type II superlattice, the "interaction" is dictated by how far the InAs wells are separated from each other, and how high the barrier is which is blocking the electrons. Fig. 10.6 describes qualitatively the broadening of the conduction band as a function of GaSb thickness.

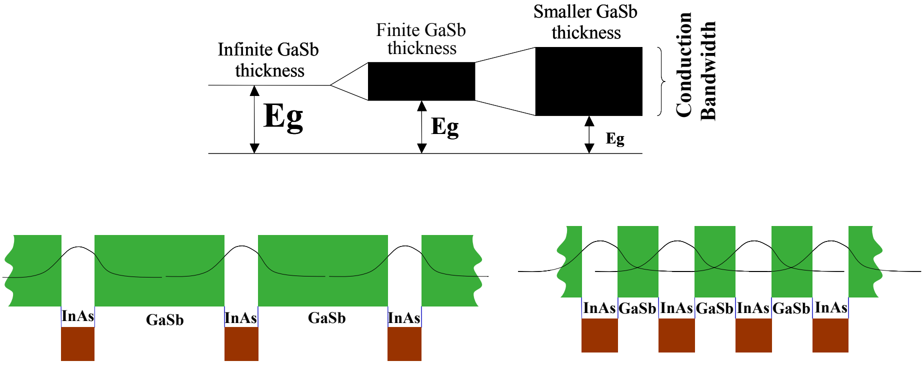


Fig. 10.6. Effect of the GaSb barrier on the conduction band level. Thin GaSb layer results to a stronger wavefunction overlap and a larger conduction band broadening.

When the GaSb layer is too thick, the "superlattice" is actually a system of non-interacting multi quantum wells. The electron energy level is solely determined by the layer thickness of the InAs wells, and electrons are strictly confined within the wells. When the GaSb thickness is reduced, the electron wavefunctions start overlapping, the electrons start seeing each other, and the single degenerated energy level is split into minibands. The broadening of the minibands pushes the lowest conduction level downward, closer to the constant valence level. Thus the thinner the GaSb barrier, the more the minibands broaden, and the smaller the energy gap becomes.

10.4.2. Quantitative calculations of electronic bandstructure

More rigorous calculations are required to precisely describe the electronic band structure of Type II superlattices. Theoretical methods that have been developed for Type II superlattices include: the k·p based theory [Grein et al. 1993] [Grein et al. 1994], the Empirical Tight Binding Model (ETBM) [Wei 2004], and the pseudo potential model [Piquini et al. 2008]. Compared to other methods, the ETBM exhibits great advantages with the capability to calculate the band structure in the whole Brillouin zone and to precisely describe the atomic layering to the superlattice, taking into account the growth imperfection. The method is also fast and does not require massive and complex numerical calculation.

The wavefunction of the superlattice is the Bloch superposition of localized atomic orbitals:

Eq. (10.1) $$\Psi_{\vec{k}}(\vec{r}) = \sum_{\vec{R}_{SL}} \sum_{\alpha} \sum_{n=1}^{N} \exp\left(i\vec{k}\cdot(\vec{R}+\vec{\tau}_n)\right) A_n^{\alpha} \varphi_n^{\alpha}(\vec{r}-\vec{R}-\vec{\tau}_n)$$

where N is the total number of atoms in one unit cell, A_n^{α} are constants, and the sum of $\vec{R}_{SL}$ runs through all the unit cells that are involved in the nearest neighbor interaction, and $\varphi_n^{\alpha}(\vec{r}-\vec{R}-\vec{\tau}_n)$ is the α orbital (α=s, p, s*) of the nth atom in the unit cell. By utilizing the formalism of the Schrödinger equation, the band structure calculation problem is reduced to an eigenvalue problem with the form:

Eq. (10.2) $$\widehat{H}\psi_k(\vec{r}) = E(k)\psi_k(\vec{r})$$

where *H* is the Hamiltonian for the system and *E(k)* is the energy eigenvalues that are dependent on $\vec{k}$ values in the 1st Brillouin zone. Using the orthogonal properties of the atomic orbitals, the eigenvalues can be rewritten:

Eq. (10.3)
$$\sum_{\vec{R}_{SL}}\sum_{\alpha}\sum_{n=1}^{N}\exp\left(i\vec{k}(\vec{R}+\vec{\tau}_n-\vec{R}'+\vec{\tau}_n')\right)\times$$
$$\times A_n^{\alpha}\iiint dr^3\varphi_m^{\beta*}(\vec{r}-\vec{R}'-\vec{\tau}_m)\hat{H}\varphi_n^{\alpha}(\vec{r}-\vec{R}-\vec{\tau}_n)=E\times A_m^{\beta}$$

The coefficient A_n^{α} and the energy $E_n(k)$ can be calculated once the Hamiltonian matrix elements are explicitly defined. In the Empirical Tight Binding Model (ETBM), the Hamiltonian matrix of a superlattice is constructed from blocks of Hamiltonian matrix for bulk semiconductors, such as GaSb, GaAs, InSb, InAs, with the same atomic sequence.

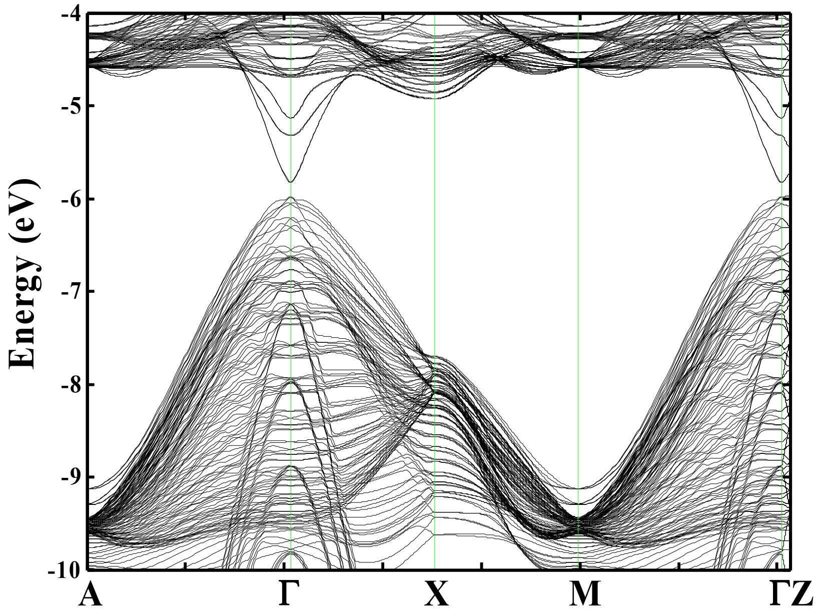


Fig. 10.7. Calculated bandstructure for a typical superlattice structure along high symmetry directions in the entire first Brillouin zone.

Fig. 10.7 sketches the band structure $E(\vec{k})$ of a typical Type II superlattice using the ETBM formalism. Energy dispersion can be calculated for hundreds of bands, from the very low valence the high conduction continuum at the GaSb's conduction band level. The minima of the conduction band and the maxima of the valence band both occur at the center of the Brillouin zone, indicating that the superlattice is a direct gap material despite the spatial separation of electron/hole in the real space. The energy gap, determined by the difference between the highest valence band

and the lowest conduction band of several superlattice designs were compared to the experimental values extracted from the optical response measurements, and showed very good agreement (Fig. 10.8).

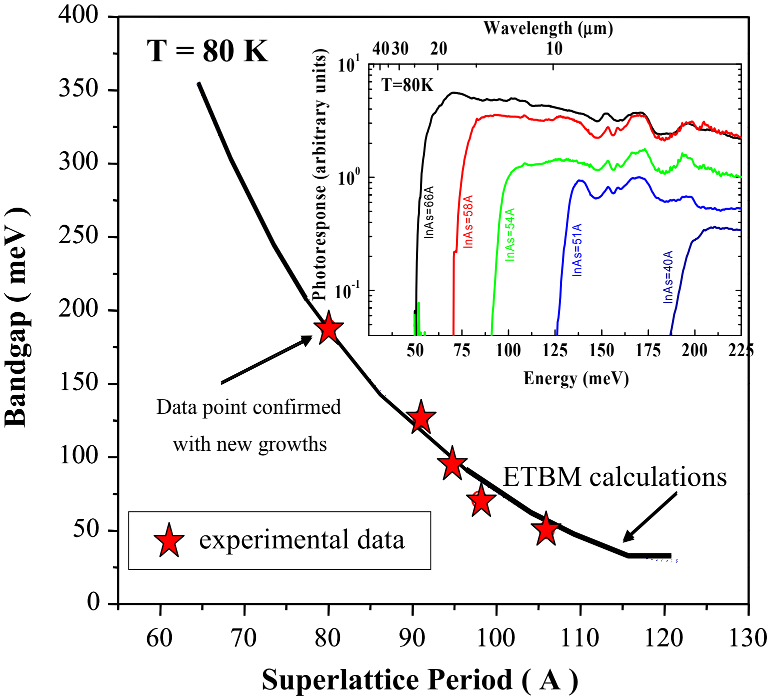


Fig. 10.8. Comparison between experimental energy gap (extracted from the optical responses in the inset) and theoretical prediction for several Type II superlattice designs.

From the band structure calculation, the conduction band and valence band limits for all possible superlattice configurations are mapped in Fig. 10.9. The evolution of the energy levels is exactly as predicted by the qualitative descriptions in the previous section. The conduction band shows a strong dependence on the InAs thickness, due to the modification of the electron's well width; and also a slight variation with the change of the GaSb layer thickness. This is due to the change of the wavefunction overlap.

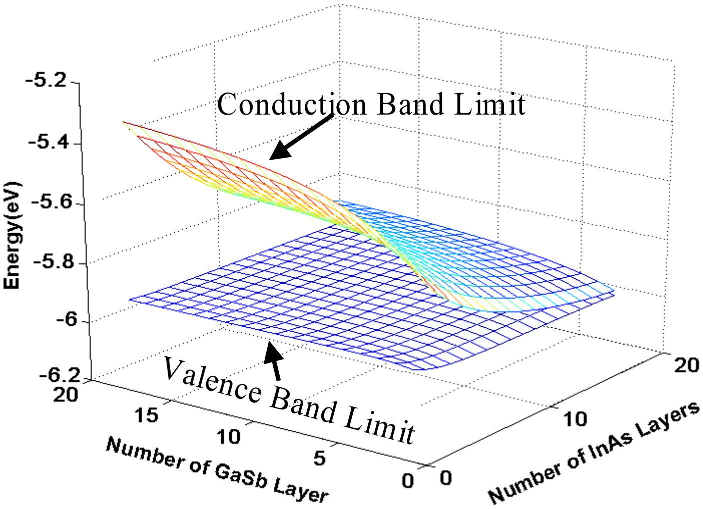


Fig. 10.9. ETBM calculation for the band edge energy distribution for Type II InAs/GaSb superlattices.

10.5. Advantages of Type II superlattice

10.5.1. Band gap engineering

Infrared photodetectors based on bulk semiconductors, such as the state-of-the-art mercury cadmium telluride or indium arsenide antimonide are often limited in the choice of the detection wavelength. For a binary semiconductor, the energy gap is not changeable. In order to adjust the cut-off wavelength, one or more elements must be added to form a ternary, quaternary compounds. However, in many cases, it is not practical for reasons of experimental growth. The compositional change in the material forces the material lattice constant to change. This can result in a strong lattice mismatch to the substrate and completely destroy the structural quality of the device.

In Type II superlattice, the energy gap is determined by the thickness of the constituent layers. Without any change in growth parameters, superlattices with cut-off wavelength ranging from 3 μm up to 32 μm have been experimentally demonstrated [Wei et al. 2004].

10.5.2. Auger suppression

In addition to the flexibility in tailoring the energy gap, there are a large number of possible superlattice configurations which allow one to generate different superlattice designs for any given cut-off wavelength. Therefore, it is possible to select the superlattice structure that can minimize non-radiative recombination, such as Auger recombination, or the Shockley-Read-Hall recombinations which use the mid-gap trap centers. Fig. 10.10 is an example of the bandgap engineering which is made to suppress Auger recombination. Three designs are selected for a cut-off wavelength around 10 μm. The GaSb quantum well for holes is thinnest in Design C, and as a consequence of quantum mechanics, the difference of the energy levels in the valence band is most pronounced in Design C. The design therefore reduces the Auger recombination process, allowing for longer carrier lifetime and better device performance.

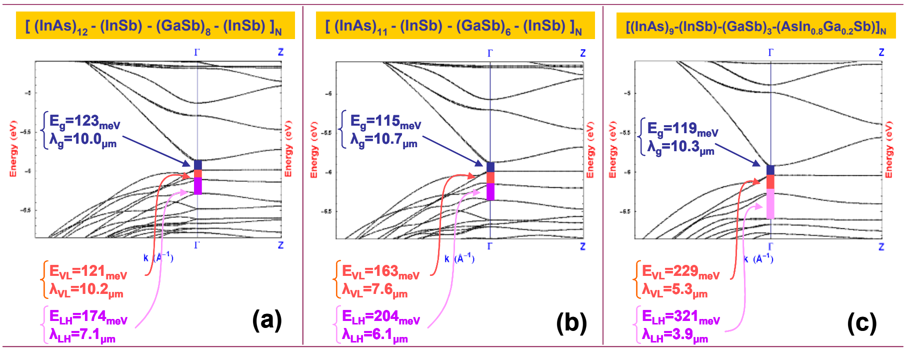


Fig. 10.10. Example of bandgap engineering for the suppression of Auger recombination. Three designs which have the same energy gap but different inter-valence band gaps. Design C is the most promising design for Auger suppression.

10.5.3. Large effective mass

Low bandgap semiconductors often produce a small effective mass. As a rule the effective mass is inversely proportional to the energy gap:

$$\text{Eq. (10.4)} \qquad m^* = \left[\frac{1}{m_e} + \frac{E_P(E_g + 2/3\Delta)}{E_g(E_g + \Delta)} \right]^{-1}$$

Where E_P is the Kane parameter, E_g is the energy gap and Δ is the spin-orbit interaction energy.

For example, the effective mass of the state-of-the-art Mercury Cadmium Telluride system is below $0.01 \cdot m_e$ when the cut-off wavelength is above 10 µm, and reduced to $0.005 \cdot m_e$ with a 15 µm cut-off (Fig. 10.11). The problem arises for LWIR and VLWIR detection. Here, the electron can tunnel through the very small forbidden gap and give rise to tunneling leakage currents. The tunneling probability depends exponentially on the square root of the effective mass; materials with lower effective mass will give rise to a stronger tunneling leakage.

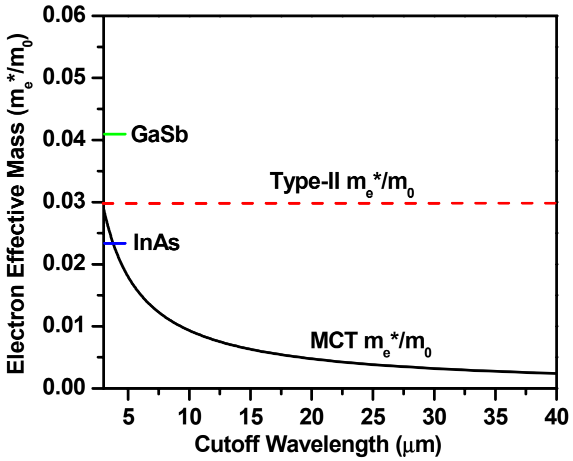


Fig. 10.11. Electron effective mass for low bandgap materials: in bulk HgCdTe, the effective mass rapidly decreases with the cut-off wavelength, while in Type II superlattice, the effective mass is similar to the effective mass of the InAs layer. Data for Type II from Fuch et al. [1997] and data for MCT from Weiler [1981].

Type II superlattices, in contrast, are predicted to have stable values of effective mass, regardless of the cut-off wavelength. This is because the effective mass is mainly determined by the overlap between electron wavefunction in adjacent InAs wells, and not by the interaction between the conduction band and the valence band like in the case of bulk materials. The large effective mass allows Type II superlattice devices to operate with lower tunneling current, and therefore, they can be optimized for higher performance.

10.5.4. Normal incident, broad band absorption

While being low dimensional and benefiting from a large effective mass, the Type II superlattice keeps the advantage of the bulk as far as the optical properties are concerned. Unlike quantum wells or quantum dots, Type II superlattices have a broad band absorption spectrum with a specific cut-off wavelength corresponding to the energy gap. Any photon with energy greater than the band gap can be absorbed and contributes to the integrated signal response. Moreover, the delocalization of the electron wavefunction allows for normal incident absorption without requiring a grating.

10.5.5. Good uniformity

Another advantage of Type II superlattices is the high quality, high uniformity and stable nature of the material. The energy gap and electronic properties of Type II superlattices are determined by the layer thicknesses rather than the molar fraction as is the case for Mercury Cadmium Telluride. The growth of Type II superlattices can be carried out with better control over the structure and with higher reproducibility. The spatial uniformity is also improved, since the effects of compositional change due to flux and temperature non-uniformity are not as important as they are in ternary/ quaternary bulk materials. Mohseni et al. has demonstrated excellent uniformity of Type II superlattices in the VLWIR regime, where in contrast, with Mercury Cadmium Telluride one experiences great difficulties in controlling the material growth (Fig. 10.12).

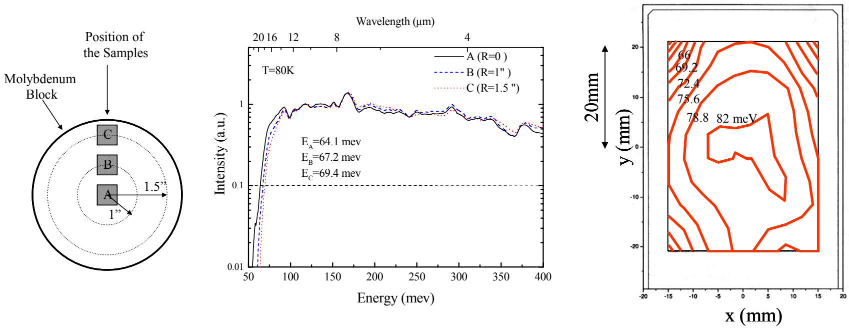


Fig. 10.12. Comparison of growth uniformity between: left & middle) Type II superlattice and right) HgCdTe. [Adapted with permission from Applied Physics Letters Vol. 77, No. 11, Mohseni, H., Tahraoui, A., Wojkowski, J., Razeghi, M., Brown, G.J., Mitchel, W.C., and Park, Y.S., “Very long wavelength infrared type-II detectors operating at 80 K,” fig. 5, pp. 1573, Copyright 2000 by American Institute of Physics and adapted with permission from, Journal of Electronic Materials Vol. 29, no. 6, Tobin, S., Hutchins, M., and Norton, P.: ‘Composition and thickness control of thin LPE HgCdTe layers using x-ray diffraction’, fig. 17 pp. 790, copyright 2000 TMS.]

10.6. Material growth and characterization

The growth of Sb-based Type II superlattice is much more difficult than the growth of bulk semiconductors or simpler heterostructures. A typical device structure consists of hundreds to thousands of sandwiched layers with an equivalent amount of interfaces. Each superlattice period requires strict control of layer thickness, and interface composition in order to achieve an overall smoothness of the whole structure. Moreover, each material has its own growth window, with a small tolerance of growth temperature and III/V flux ratio. Since the growth temperature and material flux cannot be changed within the fast switching sequence of the superlattice growth, the growth condition for the superlattice must be a compromise between the growth parameters of the constituent layers. For Type II superlattices, the growth of GaSb and InAs requires high growth temperature, while the InSb interfaces can be melted at above 450 °C. A lower growth temperature must be set in order to maintain the sharpness of the interfaces. For such low growth temperature, molecular beam epitaxy (MBE) is the best candidate to realize high quality superlattice since the growth with MOCVD will required high substrate temperature to crack the gaseous products prior to the deposition of materials. Moreover, with MBE, the layer thicknesses can be precisely determined by computer controlled shutter sequences, which can be actuated with accuracy up to 0.1 second. The Ultra High Vacuum condition of the MBE also allows for high purity and low background material growth.

High quality superlattice is often characterized by:

- Smooth surface with clear atomic steps under Atomic Force Microscopy
- Narrow peaks, High order diffraction patterns under High Resolution X-ray diffraction.
- Narrow Photoluminescence peaks at exactly the expected position.
- Sharp interface and precise layer thickness as revealed by Transmission Electron Microscopy or Scanning Tunneling Microscopy.

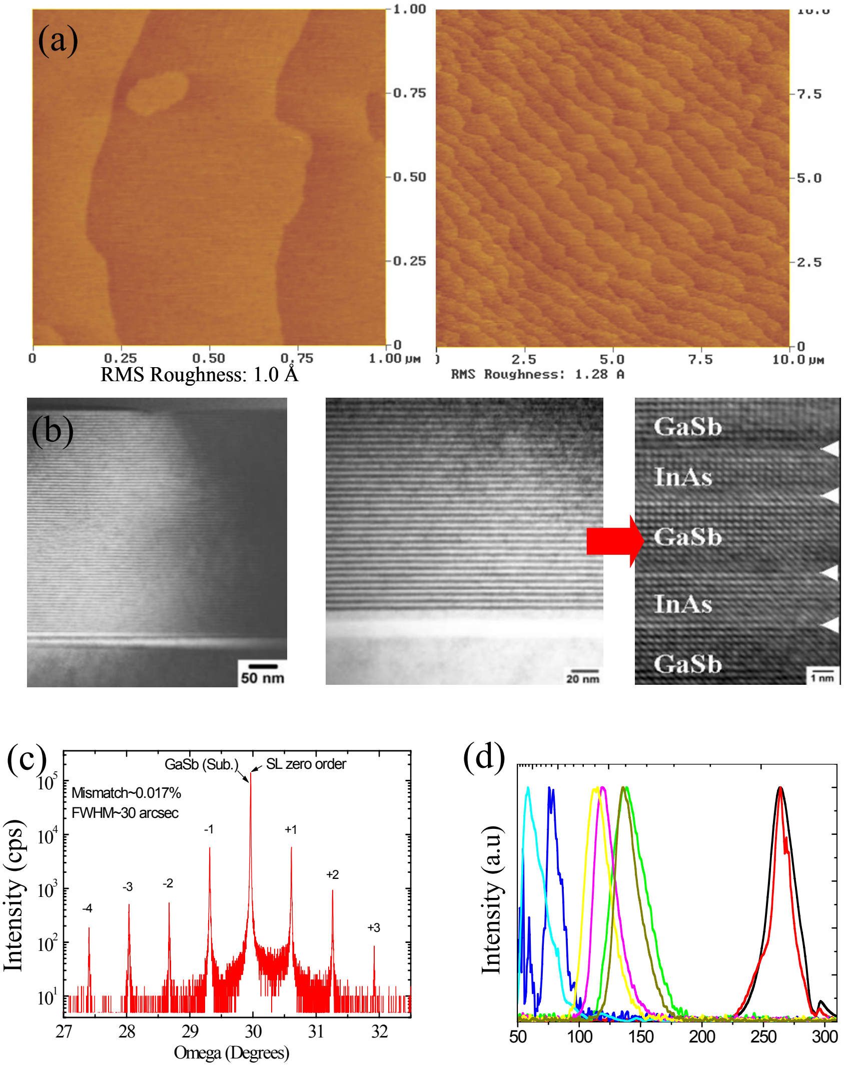


Fig. 10.13. Routine structural characterization for Type II superlattice using: a) Atomic step microscopy; b) Transmission Electron Microscopy, c) High Resolution X-ray Diffraction, d) Photoluminescence. [Parts a & d reprinted with permission from Razeghi, M., Wei, Y., Hood, A., Hoffman, D., Nguyen, B.M., Delaunay, P.Y., Michel, E., and McClintock, R., "Type-II superlattice photodetectors for MWIR to VLWIR focal plane arrays," in Infrared Technology and Applications XXXII, vol. 6206, Fig. 2a & Fig. 3 p. 62060N, copyright 2006 by SPIE, part b reprinted with permission from Razeghi, M., Nguyen, B.M., Hoffman, D., Delaunay, P.Y., Huang, E.K., Tidrow, M., and Nathan, V., "Development of material quality and structural design for high performance Type II InAs/GaSb superlattice photodiodes and focal plane arrays," in Infrared Spaceborne Remote Sensing and Instrumentation XVI, vol. 7082, Fig. 1, pp. 708204, copyright 2008 by SPIE]

10.7. Device fabrication

10.7.1. Single element device for testing

To characterize the quality and performance of the material, single element devices are fabricated using standard processing techniques presented in Chapter 3. Basic Type II device fabrication steps are presented in Fig. 10.14. Diodes with variant sizes undergo the mesa definition process with positive lithography and a combination of dry and wet etching. Then the metal contacts are defined using negative lithography and deposited using electron beam metal evaporation.

For narrow bandgap material and small size devices, the surface leakage becomes a limiting factor to the device performance and has been a known processing challenge for years. The surface leakage results from the abrupt termination of the periodic structure at the diodes' side walls. Due to the incomplete bond of surface atoms, a bending of the conduction and valence bands near the surface occurs, creating a surface conduction channel along the mesa sidewalls. Moreover, foreign adsorbents and process contaminants can further alter the surface potential and introduce trap levels within the energy gap, leading to more efficient trap-assisted tunneling currents. A good passivation layer is expected to fix the pinning of the Fermi level near the surface and to physically protect the mesa side walls from outside contaminants.

In attempts to solve the surface leakage current in small size devices, different passivation techniques using silicon based dielectrics [Nguyen et al. 2007] or polymer encapsulation [Aifer et al. 2007] [Hood et al. 2007] have brought the device results closer to actual bulk performance values of the material. Heterostructure designs with high bandgap contact regions were proposed to avoid the bending of energy bands near the device's sidewall, resulting in a significant reduction of surface leakage even on unpassivated devices. [Delaunay et al. 2007] [Vurgaftman et al. 2006]. This self-passivating property has allowed greater freedom in the processing and passivation of Type II superlattice photodiodes.

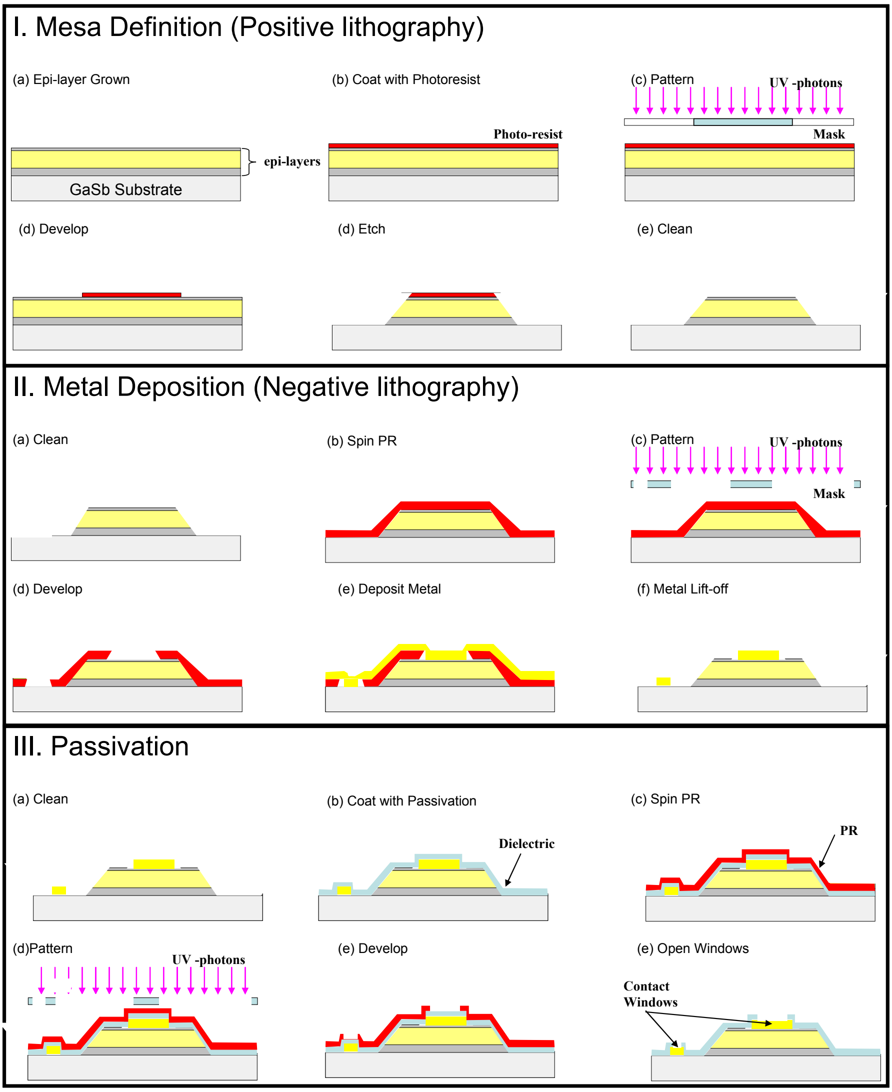

I. Mesa Definition (Positive lithography)
(a) Epi-layer Grown
(b) Coat with Photoresist
(c) Pattern
UV -photons
Photo-resist
Mask
epi-layers
GaSb Substrate
(d) Develop
(d) Etch
(e) Clean
II. Metal Deposition (Negative lithography)
(a) Clean
(b) Spin PR
(c) Pattern
UV -photons
Mask
(d) Develop
(e) Deposit Metal
(f) Metal Lift-off
III. Passivation
(a) Clean
(b) Coat with Passivation
(c) Spin PR
Dielectric
PR
(d)Pattern
UV -photons
(e) Develop
(e) Open Windows
Contact Windows

Fig. 10.14. Basic processing steps for single element device fabrication.

10.7.2. Focal plane array fabrication

The focal plane array (FPA) fabrication involves significantly more steps compared with single element detectors processing (Fig. 10.15). The FPA mesas are first formed using UV lithography and citric acid based wet etching or BCl_3 based dry etching techniques. Each FPA die has a format of 320 × 256 consisting of 25 μm × 25 μm square mesas with a pitch of 30 μm. The mesa surface is cleaned using stripper afterwards. Ti/Pt/Au is used as both top and bottom Ohmic contacts. Passivation step then gets involved to form a protection layer. Depending on different passivation techniques, different methods are applied for contact window opening. 6 µm thick indium bumps are deposited to hybridize the array to the read-out integrated circuit (ROIC, as discussed in Chapter 9, Section 9.4). Due to the oxidization nature of indium in ambient atmosphere environment, the FPA samples were diced before the indium evaporation process. The ROIC and the array are hybridized at room temperature. The edges of the gap between the ROIC and the FPA can be sealed using low viscosity epoxy before the GaSb substrate is mechanically thinned down to 60 µm. The substrate can then be completely removed using wet etching. The FPA hybrid is finally mounted on a ceramic leadless chip carrier (CLCC) with vacuum grease and then gold wire bonded for testing.

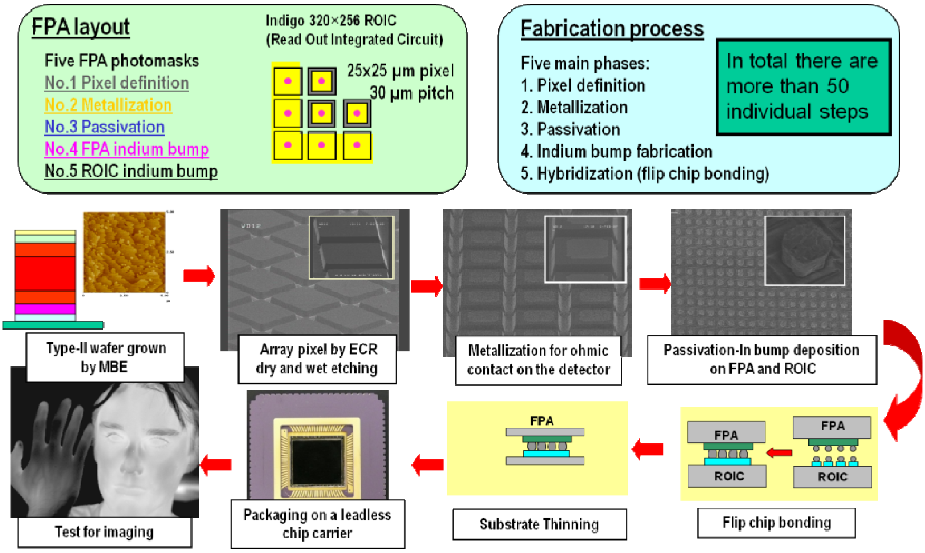

FPA layout
Indigo 320×256 ROIC
(Read Out Integrated Circuit)
Five FPA photomasks
No.1 Pixel definition
No.2 Metallization
No.3 Passivation
No.4 FPA indium bump
No.5 ROIC indium bump
25x25 µm pixel
30 µm pitch
Fabrication process
Five main phases:
1. Pixel definition
2. Metallization
3. Passivation
4. Indium bump fabrication
5. Hybridization (flip chip bonding)
In total there are more than 50 individual steps
Type-II wafer grown by MBE
Array pixel by ECR dry and wet etching
Metallization for ohmic contact on the detector
Passivation-In bump deposition on FPA and ROIC
FPA
ROIC
Flip chip bonding
Substrate Thinning
Packaging on a leadless chip carrier
Test for imaging

Fig. 10.15. Basic processing steps for the FPA fabrication.

Fig. 10.16 shows the demonstration of Type II superlattice FPA at MWIR and LWIR. MWIR FPA (Fig. 10.16a) is operable at room temperature and presents at 110 K a Noise Equivalent Temperature Difference (NEΔT) peak at 50 mK. LWIR FPA (Fig. 10.16b) presents a mean NEΔT of 26 mK at 81 K

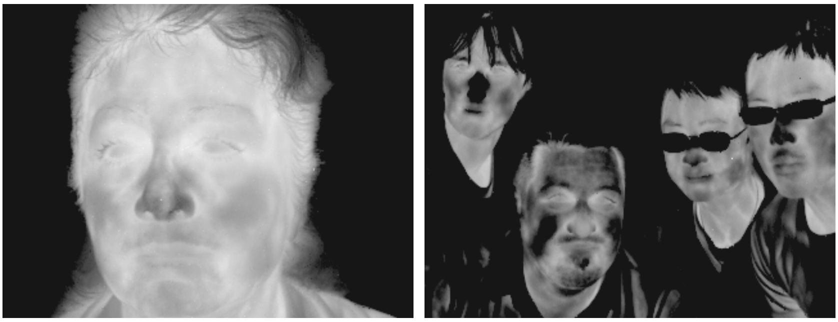


Fig. 10.16. Pictures taken with MWIR (left) and LWIR (right) cameras based on Type II superlattices.

10.8. Summary

In this chapter, we have covered the recent development of Type II superlattices for infrared detection and imaging. The concepts and properties of Type II superlattices were discussed, illustrating how quantum mechanics can be applied to this material. In addition we described the advantages of the material system as compared to the state-of-the-art Mercury Cadmium Telluride, and we explained why Type II superlattices were considered as the material of choice for the future of infrared detection. The fabrication processes of infrared detectors were also presented, with great detail.

References

Aifer, E.H., Warner, J.H., Stine, R.R., Vurgaftman, I., Canedy, C.L., Jackson, E.M., Tischler, J.G., Meyer, J.R., Petrovykh, D.Y., and Whitman, L.J., "Passivation of W-structured type-II superlattice long-wave infrared photodiodes," *Proceedings of the SPIE* 6542, pp. 654203-654212, 2007.

Cabanski, W.A., Eberhardt, K., Rode, W., Wendler, J.C., Ziegler, J., Fleissner, J., Fuchs, F., Rehm, R.H., Schmitz, J., Schneider, H., and Walther, M., "Third-generation focal plane array IR detection modules and applications," *Proceedings of the SPIE* 5406, pp. 184-192, 2004.

Delaunay, P.-Y., Hood, A., Nguyen, B.M., Hoffman, D., Wei, Y., and Razeghi, M., “Passivation of type-II InAs/GaSb double heterostructure,” *Applied Physics Letters* 91(9), pp. 091112-091113, 2007

Delaunay, P.-Y., and Razeghi, M., "High-performance focal plane array based on type-II InAs/GaSb superlattice heterostructures," *Proceedings of the SPIE* 6900, pp. 69000M-69010, 2008.

Fuchs, F., Weimer, U., Pletschen, W., Schmitz, J., Ahlswede, E., Walther, M., Wagner, J., and Koidl, P., "High performance InAs/$Ga_{1-x}In_xSb$ superlattice infrared photodiodes," *Applied Physics Letters* 71(22), pp. 3251-3253, 1997.

Grein, C., Young, P., Ehrenreich, H., and McGill, T., "Auger lifetimes in ideal InGaSb/InAs superlattices," *Journal of Electronic Materials* 22(8), pp. 1093-1096, 1993.

Grein, C.H., Cruz, H., Flatte, M.E., and Ehrenreich, H., "Theoretical performance of very long wavelength InAs/$In_xGa_{1-x}Sb$ superlattice based infrared detectors," *Applied Physics Letters* 65(20), pp. 2530-2532, 1994.

Hood, A., Delaunay, P.Y., Hoffman, D., Nguyen, B.-M., Wei, Y., Razeghi, M., and Nathan, V., "Near bulk-limited R_0A of long-wavelength infrared type-II InAs/GaSb superlattice photodiodes with polyimide surface passivation," *Applied Physics Letters* 90(23), pp. 233513-233513, 2007.

Johnson, J.L., Samoska, L.A., Gossard, A.C., Merz, J.L., Jack, M.D., Chapman, G.R., Baumgratz, B.A., Kosai, K., and Johnson, S.M., "Electrical and optical properties of infrared photodiodes using the InAs/$Ga_{1-x}In_xSb$ superlattice in heterojunctions with GaSb," *Journal of Applied Physics* 80(2), pp. 1116-1127, 1996.

Meyer, J.R., Hoffman, C.A., Bartoli, F.J., and Ram-Mohan, L.R., "Type-II quantum-well lasers for the mid-wavelength infrared," *Applied Physics Letters* 67(6), pp. 757-759, 1995.

Mohseni, H., Michel, E., Sandoen, J., Razeghi, M., Mitchel, W., and Brown, G., "Growth and characterization of InAs/GaSb photoconductors for long wavelength infrared range," *Applied Physics Letters* 71(10), pp. 1403-1405, 1997.

Nguyen, B.M., Hoffman, D., Delaunay, P.-Y., and Razeghi, M., "Dark current suppression in type II InAs/GaSb superlattice long wavelength infrared photodiodes with M-structure barrier," *Applied Physics Letters* 91(16), pp. 163511-163513, 2007.

Nguyen, B.M., Hoffman, D., Wei, Y., Delaunay, P.-Y., Hood, A., and Razeghi, M., "Very high quantum efficiency in type-II InAs/GaSb superlattice photodiode with cutoff of 12 μm," *Applied Physics Letters* 90(23), pp. 231108-231103, 2007.

Nguyen, B.M., Hoffman, D., Huang, E.K.-W., Delaunay, P.-Y., and Razeghi, M., "Background Limited Long wavelength infrared Type-II InAs/GaSb Superlattice Photodiodes operating at 110 K," *Applied Physics Letters* 93, pp. 123502-1, 2008.

Nguyen, B.M., Razeghi, M., Nathan, V., and Brown, G.J., "Type-II M structure photodiodes: an alternative material design for mid-wave to long wavelength infrared regimes," *Proceedings of the SPIE* 6479, pp. 64790S-64710, 2007.

Nguyen, J., and Razeghi, M., "Techniques for high quality SiO_2 films," *Proceedings of the SPIE* 6479, pp. 64791K-64798K, 2007.

Piquini, P., Zunger, A., and Magri, R., "Pseudopotential calculations of band gaps and band edges of short-period $(InAs)_n/(GaSb)_m$ superlattices with different substrates, layer orientations, and interfacial bonds," *Physical Review B (Condensed Matter and Materials Physics)* 77(11), pp. 115314-115316, 2008.

Rodriguez, J.B., Plis, E., Bishop, G., Sharma, Y.D., Kim, H., Dawson, L.R., and Krishna, S., "nBn structure based on InAs/GaSb type-II strained layer superlattices," *Applied Physics Letters* 91(4), pp. 043514-043512, 2007.

Sai-Halasz, G.A., Tsu, R., and Esaki, L., "A new semiconductor superlattice," *Applied Physics Letters* 30(12), pp. 651-653, 1977.

Smith, D.L., and Mailhiot, C., "Proposal for strained Type II superlattice infrared detectors," *Journal of Applied Physics* 62(6), pp. 2545-2548, 1987.

Tsu, R., Chang, L.L., Sai-Halasz, G.A., and Esaki, L., "Effects of Quantum States on the Photocurrent in a 'Superlattice'," *Physical Review Letters* 34(24), pp. 1509, 1974.

Vurgaftman, I., Aifer, E.H., Canedy, C.L., Tischler, J.G., Meyer, J.R., Warner, J.H., Jackson, E.M., Hildebrandt, G., and Sullivan, G.J., "Graded band gap for dark-current suppression in long-wave infrared W-structured Type-II superlattice photodiodes," *Applied Physics Letters* 89(12), pp. 121114-121113, 2006.

Walther, M., Schmitz, J., Rehm, R., Kopta, S., Fuchs, F., Fleibner, J., Cabanski, W., Ziegler, J., "Growth of InAs/GaSb short-period superlattices for high-resolution mid-wavelength infrared focal plane array detectors," Journal of Crystal Growth 278(1-4), pp. 156-161, 1995.

Walther, M., Rehm, R., Fleissner, J., Schmitz, J., Ziegler, J., Cabanski, W., and Breiter, R., "InAs/GaSb type-II short-period superlattices for advanced single and dual-color focal plane arrays," *Proceedings of the SPIE* 6542, pp. 654206-654208, 2007.

Wei, Y., and Razeghi, M., "Modeling of type-II InAs/GaSb superlattices using an empirical tight-binding method and interface engineering," *Physical Review B (Condensed Matter and Materials Physics)* 69(8), pp. 085316-085317, 2004.

Further reading

Dereniak, E.L., and Boreman, G.D., *Infrared Detectors and Systems*, John Wiley & Sons, New York, 1996.

Henini, M., and Razeghi, M., *Handbook of infra-red detection technologies* Elsevier Science Ltd., 2002.

Hudson Jr., R.D., *Infrared System Engineering*, John Wiley & Sons, New York, 1969.

Rogalski, A., *Infrared Photon Detectors*, Bellingham, Washington, 1995.

Problems

1. A superlattice design consists of 13 Monolayers of InAs, 7 Monolayers of GaSb and the two interfaces are purely InSb. Calculate the lattice mismatch of the design to GaSb substrates.
   a_{InAs}=6.0584 Å, a_{GaSb} = 6.0959 Å, a_{InSb}=6.4794 Å.

2. The Type II superlattice design with 13 Monolayers of InAs and 7 Monolayers of GaSb has a cut-off wavelength of 11 μm. Given that at this wavelength, the change of one monolayer of either InAs or GaSb

corresponds to a 1 μm shift in the cut-off wavelength. Propose two superlattice designs for a cut-off wavelength of 10 μm.

3. In Fig. 10.14, on page 388 what kind of lithography was used in the passivation steps (positive or negative lithography)? Design the mask for the other kind of lithography, and discuss the advantages/ disadvantages of both.

11. Quantum Dot Infrared Photodetectors

11.1. Introduction
- 11.1.1. Operating principles of QWIPs and QDIPs
- 11.1.2. Photocurrent
- 11.1.3. Dark current
- 11.1.4. Noise

11.2. Advantages of QDIPs
- 11.2.1. Introduction
- 11.2.2. High gain and the phonon bottleneck
- 11.2.3. Low dark current
- 11.2.4. Normal incidence absorption
- 11.2.5. Versatility
- 11.2.6. Summary

11.3. Quantum dot fabrication for QDIPs
- 11.3.1. Introduction
- 11.3.2. The formation of quantum dots in the Stranski-Krastanov growth mode
- 11.3.3. Properties of Stranski-Krastanov grown dots and their effect on QDIP performance
- 11.3.4. Quantum dot size
- 11.3.5. Quantum dot shape
- 11.3.6. Quantum dot density
- 11.3.7. Quantum dot uniformity
- 11.3.8. Conclusion and future directions for dot fabrication

11.4. Review of actual QDIP performance
- 11.4.1. Introduction
- 11.4.2. High operating temperature
- 11.4.3. FPA imaging
- 11.4.4. Summary

11.5. Summary

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_11,
© Springer Science+Business Media, LLC 2010

11.1. Introduction

One area of optoelectronics where low dimensional systems have been applied with a great deal of success is infrared detection. The quantum well infrared photodetector (QWIP) is one such application. As its name implies, the QWIP is based on a 2D system or, equivalently, confinement in one dimension. The next natural evolution of this detector design would be the use of quantum wire 1D systems and then quantum dot (QD) 0D systems. (The fundamentals of each low dimensional system are covered in elsewhere in this text.) In practice, however, detector evolution did not go from well to wire to dot because growth and fabrication technologies have made it easier to create dots instead of wires. Quantum dot infrared photodetectors or QDIPs have become a major research interest as a result. The harnessing of quantum effects and nanotechnology for improved IR detectors is a not an easy task, but, as this chapter will show, it could be a very fruitful undertaking.

This chapter begins by introducing the basic intersubband detector operating principles, which are shared by QWIPs and QDIPs. Then we see where QDIP operation deviates from these simple principles when we discuss the theoretical advantages of QDIPs. Next we look at the growth technology for making the QDs that go into the QDIPs. The capabilities and limitations of the growth technology directly relate to whether or not the theoretical advantages of QDIPs can be achieved. Finally, we finish by reviewing some of the major accomplishments in QDIP technology to date.

11.1.1. Operating principles of QWIPs and QDIPs

Before we can explore the advantages of quantum dots (QDs), it is necessary to first review the basic operating mechanisms of an intersubband (ISB) detector. In this section we will cover operating principles that generally apply to both QWIPs and QDIPs. The differences that make the QDIP a more promising detector technology will be covered in subsection 11.2.

ISB devices are unipolar devices involving only electrons or holes. Most ISB detectors, particularly QDIPs, are made *n*-type (to take advantage of the higher electron mobility) and so the diagrams and discussions in this chapter assume electrons in the conduction band. While *p*-type devices behave similarly to *n*-type devices they have interesting differentiating features which are outside of the context of the discussion here. Basic QWIPs and QDIPs have the structures shown in Fig. 11.1. The structures are similar with the exception that QDIPs have QDs in place of the quantum well (QW). A schematic of the bandstructure of a QWIP or QDIP is shown in Fig. 11.2. In interpreting Fig. 11.2 it is important to remember that the dimensionality of the electron confining potential is not illustrated since there are not enough physics dimensions to easily illustrate 3D confinement in such a band diagram. This is why we can say the picture applies to either a QWIP or QDIP. With these two pictures in mind let us now consider the basic processes in these detectors.

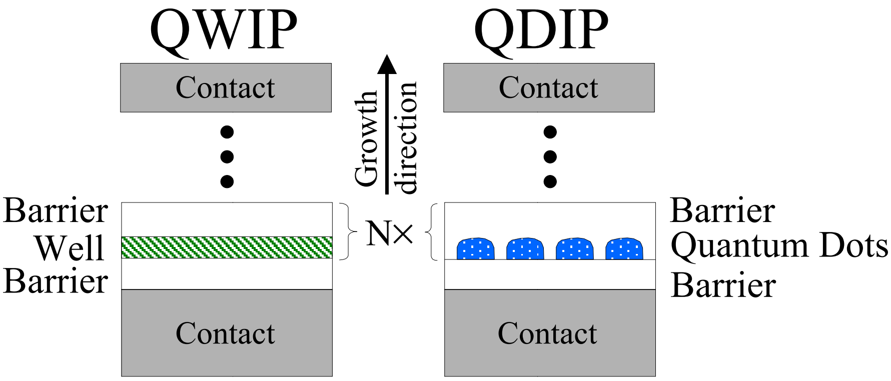

QWIP
QDIP
Contact
Contact
Growth direction
Barrier
Well
Barrier
N×
Barrier
Quantum Dots
Barrier
Contact
Contact

Fig. 11.1. Schematic device structure of a QWIP (right) and a QDIP (left).

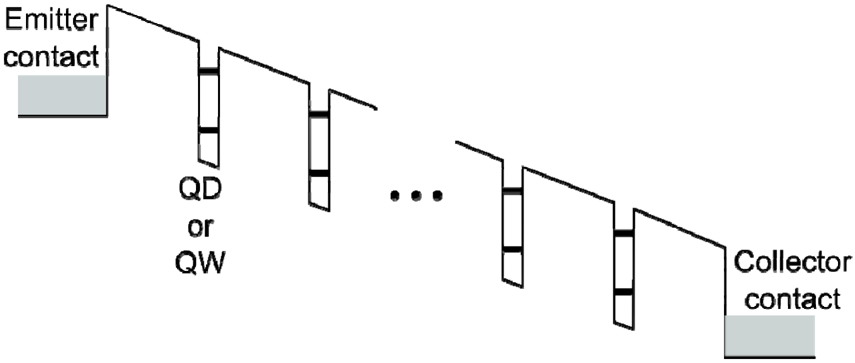

Emitter contact
QD or QW
Collector contact

Fig. 11.2. Bandstructure of a QWIP or QDIP under bias

11.1.2. Photocurrent

We will start the discussion by looking at where the signal comes from in a QWIP or QDIP. Photon detection occurs when incident infrared light excites an electron in the ground state out of the QW or QD and into the continuum. There are two possible paths for the photoexcited electron to escape out of the confining potential, which are depicted in Fig. 11.3.

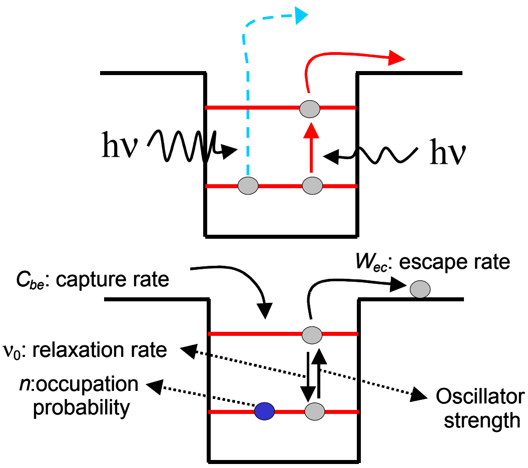


Fig. 11.3. (Top) Schematic of the photoexcitation mechanisms in an intersubband detector (Bottom) Schematic of the factors that affect the photoexcitation process

In the first path (dashed) the light directly excites the electron from the ground state to the continuum. In the second path (solid), the light first excites the electron from the ground state to an excited state. Then, from the excited state the photoexcited carrier can thermally escape to or tunnel to the continuum. While under an applied bias, once an electron has been excited to the continuum it will be swept towards the contacts and contribute to the photocurrent.

Also depicted in Fig. 11.3 are the relaxation or recapture paths, which occur when a photocurrent electron does not make it out of the well or dot or does not make it to a contact. While the photoexcitation process just described is caused by photons, these relaxation processes are caused by phonons. The time it takes for a carrier traveling in the continuum to be recaptured into a well or dot is the carrier lifetime. The ratio of the carrier lifetime to the time it takes for an electron to travel across the entire device from contact to contact is defined as the gain g. Gains greater than unity means that a photogenerated electron can travel through the device more than once creating a greater signal per photoexcitation event which increases the responsivity.

Eq. (11.1) $$g = \frac{\tau_{lifetime}}{\tau_{transit}} = \frac{\text{carrier lifetime}}{\text{carrier transit time}}$$

11.1.3. Dark current

The dark current is a non-signal current flow in the detector or current that flows without light. In imaging applications low dark current is desired because the amount of current that can be read at one time is fixed, and it is preferable for as much of that fixed amount of current to be photocurrent not dark current. In an intersubband detector there are three primary dark current mechanisms, as illustrated in Fig. 11.4. They are:

A. Thermionic emission – Electrons in the ground state are thermally excited to the continuum
B. Thermally assisted tunneling – Electrons are thermally excited to an excited states and then tunnel out of the excited state into the continuum
C. Sequential tunneling – Electrons tunnel directly between well or dot ground state and eventually to the contact

Mechanism A and B are the dominant processes for most of the devices and operating conditions relevant to our discussion. Mechanisms A and B are most strongly affected by the energy level structure of the dot or well, the operating temperature, and the applied bias.

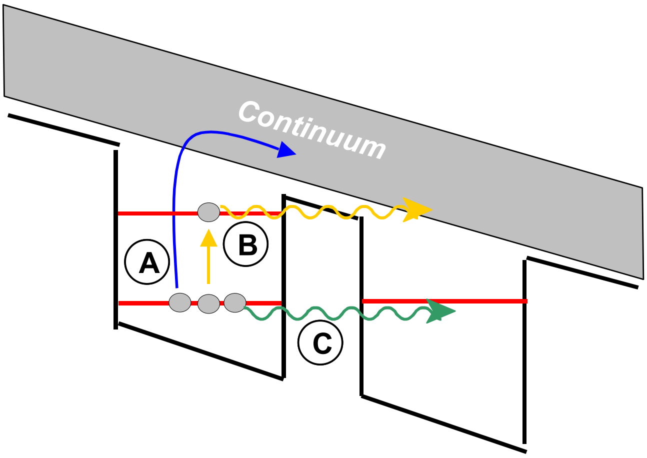


Fig. 11.4. Dark current mechanisms in QWIPs and QDIPs. (A) Thermionic emission (B) Thermally assisted tunneling (C) Sequential tunneling

11.1.4. Noise

For typical devices and operating conditions in QWIPs and QDIPs, generation-recombination (GR) noise is the dominant noise source. The GR noise is determined by the following equation, Eq. (11.2) $i_{GR}^2 = 4eg\Delta fI_D$

where i_{GR} is the GR noise, e is the electron charge, g is the gain, Δf is the noise bandwidth, and I_D is the dark current. GR noise is a result of the statistical fluctuation in the generation and recombination rates of electrons between different energy states in the material.

11.2. Advantages of QDIPs

11.2.1. Introduction

In the mid-1990s, Ryzhii published the first detailed analysis of the benefits of using quantum dots for infrared detection [Ryzhii et al. 1996]. Since then, quantum dot infrared photodetectors (QDIPs) have been consistently studied as a candidate for next generation of infrared detectors. Later, Phillips and Martyniuk et al. further developed Ryzhii's original model of QDIP performance and showed that QDIPs have the potential for excellent IR detector performance [Martyniuk et al. 2008][Phillips et al. 2002]. The interest in QDIPs comes primarily from their predicted ability to achieve high performance at high operating temperatures (near room temperature). High temperature operation is an important but difficult technological hurdle for current photon detectors, which typically require some level of cryogenic cooling. The high operating temperature capability of QDIPs comes from two quantum dot-related effects: low dark current and high photoelectric gain. QDIPs also have two other important technological benefits which we will also discuss here.

11.2.2. High gain and the phonon bottleneck

From the discussion on the operation of intersubband detectors, we can see that carrier relaxation is an important factor in detector performance, and the dominant relaxation paths in intersubband devices are phonon-mediated ones. Early theoretical studies of low dimensional structures predicted a phenomenon called the *phonon bottleneck*, where the delta function-like density of states significantly slows down carrier transitions between energy levels in a QD system. This is illustrated for carrier relaxation in Fig. 11.5. For the 1D and 2D cases there is a continuum of states in at least one dimension, whereas in the 3D system there is only the fully quantized, delta function-like ground state. As a result, phonon mediated processes are less likely to occur since the phonon energy must exactly match the energy spacing between the QD levels. Additionally, in a typical QD system, the energy level spacing for QDs with only 1 or 2 levels ends up being larger than the typical LO phonon energy so there is little opportunity for phonon scattering. Since the phonon-mediated relaxation paths are blocked the excited state carrier lifetimes should lengthen from the 1–10 ps measured in QW systems to the nanosecond range predicted for QD systems.

In a detector context, this means that photoexcited carriers will have a longer lifetime both in excited states and in the continuum. Recall that gain is defined as $\frac{\tau_{\text{lifetime}}}{\tau_{\text{transit}}}$ in Eq. (11.1). So a longer lifetime gives higher gain, which then yields higher responsivity, since responsivity proportional to gain.

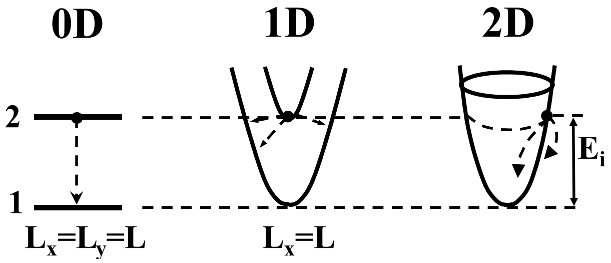


Fig. 11.5. Illustration of the differences in phonon mediated electron relaxation in 0D, 1D, and 2D structures. In 1D and 2D there are several allowed phonon relaxation transitions of an electron at energy E_i, whereas in 0D, as soon as E_i is greater than the LO phonon energy, no single phonon transitions are allowed."

11.2.3. Low dark current

The other primary advantage for high operating temperature is the expectation that QDIPs will have low dark current. The dominant dark current mechanism in intersubband devices is typically thermionic emission, when an electron is thermally excited out of the well. Other possible mechanisms are tunneling-assisted thermionic emission and dot-to-dot tunneling. With 3D confinement and the resultant well-defined quantized state, the thermionic emission path and the photoexcitation path become one and the same, competing with each other. Photoexcitation is typically faster than thermionic emission, causing a reduction in dark current. Also, since there are only the 1 or 2 quantized states in the QD, any thermal excitation must take the electron out of the QD in only one step, making the activation energy higher for QD systems.

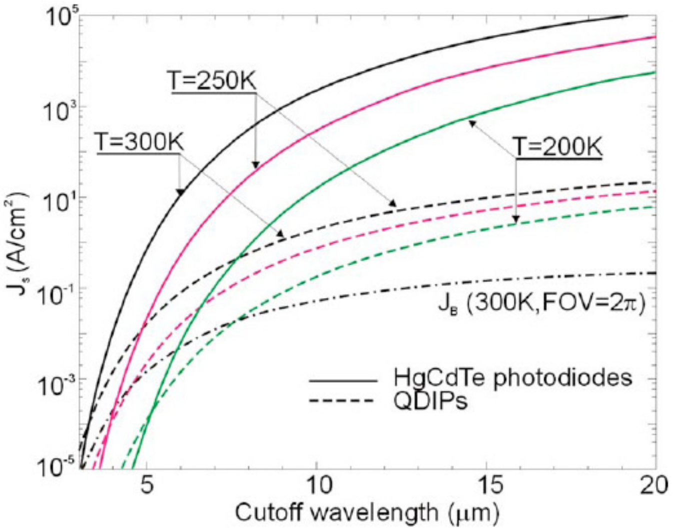


Fig. 11.6. Calculated dark current for QDIPs compared to HgCdTe as a function of wavelength. [Reproduced with permission from Journal of Applied Physics Vol.104, P. Martyniuk, S. Krishna, and A. Rogalski1, "Assessment of quantum dot infrared photodetectors for high temperature operation," fig. 2, p. 034314-4, Copyright 2008, American Institute of Physics (AIP).]

11.2.4. Normal incidence absorption

In comparison to QWIPs, QDIPs have the advantageous capability to absorb normally incident light. Most detector applications utilize a normal incidence light configuration where the incoming light signal is traveling perpendicular to the growth plane of the detector elements and is therefore randomly polarized in the growth plane. This is illustrated in Fig. 11.7. The strength of the absorption of the incident light is related to the oscillator strength, which can be written as,

Eq. (11.3)

$$f_{12} \propto \left| \left\langle F_1 \middle| \vec{\varepsilon} \cdot \vec{p} \middle| F_2 \right\rangle \right|^2 \propto \left| \left\langle F_1 \middle| \left[\varepsilon_x (\partial / \partial x) + \varepsilon_y (\partial / \partial y) + \varepsilon_z (\partial / \partial z) \right] \middle| F_2 \right\rangle \right|^2$$

where F_1 and F_2 are the electron envelope functions, ε is the polarization vector for the incident infrared light and the p is the momentum operator. For normally incident light, ε_x and ε_y are nonzero and ε_z is zero. In the QW case, F_1 and F_2 are only functions of z (the confined dimension). So the for oscillator strength, in the x and y directions the partial derivative goes to zero and in the z direction ε_z is zero, therefore f_{12} is zero. In the QD case the, F_1 and F_2 are functions of x, y, and z, therefore even though ε_z is zero, the first two terms are nonzero and contribute to the oscillator strength for normal incidence.

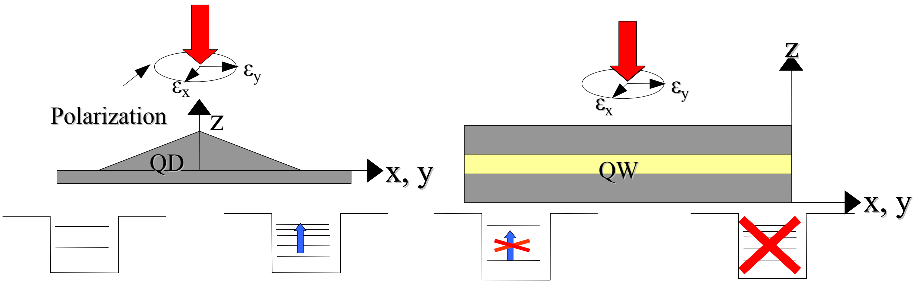


Fig. 11.7. Schematic of the polarization sensitivity of QD and QW absorption

In QWIPs the normal incidence limitation is dealt with by placing a grating on the detector surface that redirects the light, changing the polarization orientation. In QDIPs with normal incidence absorption this extra, non-trivial processing step could be removed, simplifying the device processing. Also the polarization sensitivity limits the ultimate quantum efficiency of a QWIP to 50%, whereas an ideal QDIP would not have this limit on the quantum efficiency.

11.2.5. Versatility

This last advantage of QDIPs is not a performance advantage but a design or technology advantage. Compared with bulk and QW detectors, QDIPs have more numerous adjustable design parameters and thus greater potential versatility. For example, to tune the wavelength, the main adjustment parameter for bulk systems is material composition. In a QW system the parameters are material composition and well thickness, with strain being a secondary parameter. Finally, in a QD system material composition, strain, and QD size and shape can all be used to tune the wavelength. In QD systems there are more available "knobs to turn" to achieve the desired outcome. However, this can be a double-edged sword if the parameters are difficult to control – which is, in fact, the case for QDs as discussed in subsection 11.3 which covers QD growth.

11.2.6. Summary

The above advantages if fully realized make QDIPs a very promising technology for infrared imaging applications. This can be seen in the calculations that take these advantages into account by Martyniuk et al. [2008] in Fig. 11.8 where the ultimate QDIP detectivity is shown to be comparable to or better than HgCdTe detectors particularly at room temperature and for long wavelength detection.

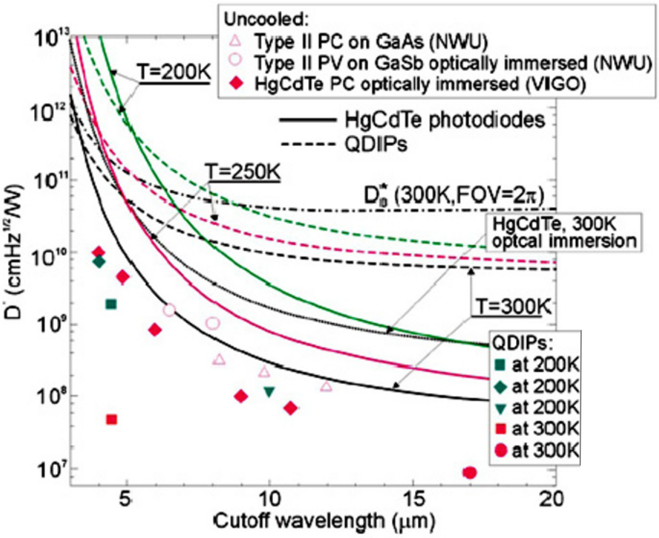


Fig. 11.8. The lines show theoretical D comparisons for HgCdTe and QDIP detectors. The points show various experimental data from high operating temperature detectors. [Reproduced with permission from Journal of Applied Physics Vol.104, P. Martyniuk, S. Krishna, and A. Rogalski1, "Assessment of quantum dot infrared photodetectors for high temperature operation," fig. 3, p. 034314-5, Copyright 2008, American Institute of Physics (AIP).]*

11.3. Quantum dot fabrication for QDIPs

11.3.1. Introduction

The biggest hurdle in the study of low dimensional semiconductor systems has always been the physical realization of such systems. Since the relevant features sizes for these systems are at the nanometer level, even slight imperfections in the material can mask or eliminate any quantum size effects. For example, in early experimental investigations of quantum well (QW) systems, low quality well material and rough well-barrier interfaces prevented the observation of quantum size effects. These problems were overcome with the advent of MBE and MOCVD epitaxy technologies. After much progress was made in QWs, attention turned towards structures with lower dimensionality, but quantum dots (QDs) faced a similar fabrication challenges.

For detector applications, QD systems should have the following qualities:

1. QDs of appropriate size
2. Uniform QDs in high density arrays
3. Defect-free QDs

Creating quantum dots with these qualities proved difficult by methods such as direct patterning using electron beam lithography. In QD systems, the key innovation was the rediscovery and application of the Stranski-Krastanov (SK) growth mode, which is a 3D crystal growth mode that can occur in the epitaxy of III-V materials. Of the available methods for creating quantum dots, the use of the SK growth mode comes closest to meeting the three main requirements for using quantum dots in detectors. An overview of the different epitaxial growth modes is presented in Chapter 1. This section looks more closely at the SK growth mode which is also often referred to as quantum dot fabrication by self-assembly.

11.3.2. The formation of quantum dots in the Stranski-Krastanov growth mode

The Stranski-Krastanov (SK) growth mode occurs when growing lattice mismatched materials where the QD material is grown on a substrate or matrix layer with smaller lattice constant. The first few deposited layers of the QD material grow in a flat, layer-by-layer fashion. This flat layer is called the wetting layer. Since the QD material is lattice mismatched, the wetting layer is pseudomorphically strained and the strain builds up with increasing material deposition. Beyond a certain critical thickness the wetting layer spontaneously reorganizes and continued deposition of material results in the growth of 3D dot features on top of the thin wetting layer. This process is illustrated in Fig. 11.9. SK growth typically occurs when the lattice mismatch is around 3 to 10% (though these limits are not well-defined). The most studied SK QD growth in III-V materials is In(Ga)As on GaAs. Other systems that have been studied include InAs on InP, Sb-based III-Vs, and III-Nitrides.

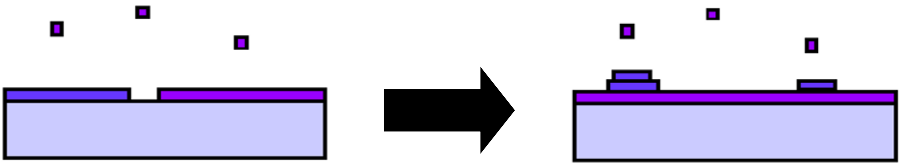


Fig. 11.9. Illustration of the Stranski-Krastanov growth mode.

This growth process can be understood by considering the interplay between the surface, interface, and strain energies in this situation. Initially the sum of the epilayer surface energy and the interface energy is lower than the surface energy of the matrix, therefore it is favorable to have layer-by-layer growth of the QD material. After several monolayers of deposition, the buildup of strain energy makes it no longer favorable to have flat, layer-by-layer growth. With this built-up energy the material needs to relieve the strain and does so by forming 3D dot structures.

The key factor here for optoelectronic devices is that the relaxation process in SK growth can be controlled to produce defect-free (also called coherent) QDs. This coherent relaxation was first noted in the mid-80's [Goldstein et al. 1985]. Since the dot material is experiencing a compressive strain, when the coherent relaxation takes place the lattice constant expands towards the dot edges, as illustrated in Fig. 11.10. This coherent relaxation means that the QDs grown by SK mode do not have the defects that reduce the 3D confinement in other QD fabrication methods.

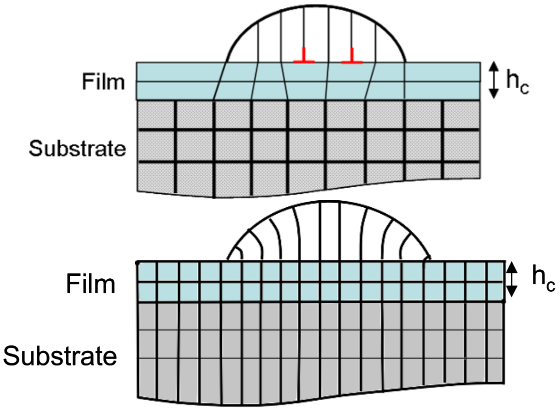


Fig. 11.10. (Left) Schematic illustrating an SK dot that has relaxed via defect formation and (Right) schematic of a dot that has coherent relaxed without defects.

11.3.3. Properties of Stranski-Krastanov grown dots and their effect on QDIP performance

The most important QD parameters are: size, shape, density, and uniformity. We will discuss the significance of each parameter for detector performance and the typically achieved growth result from Stranski-Krastanov (SK) growth. Fig. 11.11 shows a representative example of quantum dots grown by the SK growth mode.

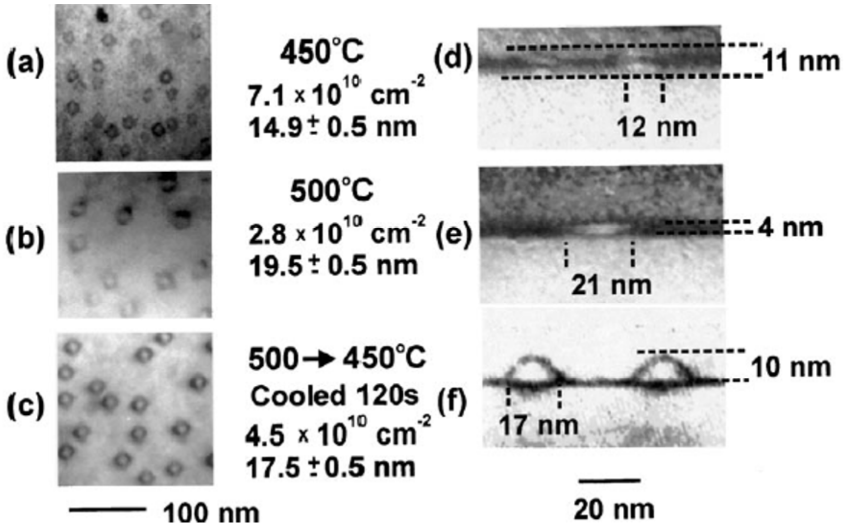


Fig. 11.11. Example of InAs/GaAs quantum dots grown by SK mode in MBE. (a), (b), and (c) are plan view TEM images. (d),(e), and (f) are cross sectional TEM images. The labels indicate the growth conditions, dot densities and dimensions. [Reproduced with permission from Semiconductor Science and Technology Vol.16, N.N. Ledentsov1, V.A. Shchukin1, D. Bimberg, V.M. Ustinov, N.A. Cherkashin, Y.G. Musikhin, B.V. Volovik, G.E. Cirlin, and Z.I Alferov, "Reversibility of the island shape, volume and density in Stranski-Krastanov growth," fig. 2, p. 505, Copyright 2001, IOP Publishing Ltd.]

11.3.4. Quantum dot size

For QDs to be useful they must be of an appropriate size. The absolute minimum size for a dot is the smallest size for which there will still be one electron energy level in the dot. For a spherical dot this size is:

Eq. (11.4) $$D_{\min} = \frac{\pi\hbar}{\sqrt{\left(2m_e^*\Delta E_c\right)}}$$

In practice the dot must actually be larger than this minimum size in order to provide strong carrier localization at non-zero temperature. This is easily achievable in SK growth. Instead the challenge is in not surpassing the maximum useful dot size. The maximum size has been surpassed when the QDs have more than 1 or 2 energy levels within the dot and no longer show strong three-dimensional confinement effects. When there are more than two energy levels in a dot, these additional levels can break the phonon bottleneck and dark current reduction. Also if the dot is too large the energy levels will become more widely spread further reducing the phonon bottleneck. QDIP device models showing good performance have been based on dot sizes around 15 nm [Martyniuk et al. 2008][Phillips et al. 2002]. In the literature QDs grown by SK self assembly vary from about 10–50 nm for coherent dots.

Size control of the quantum dots is also important to control the position of the excited state energy level of the QD relative to the continuum. If the excited state is too deep in the dot a high bias may be required to extract a photoexcited electron from a QD. This is the case for the calculated dot configurations shown in Fig. 11.12.

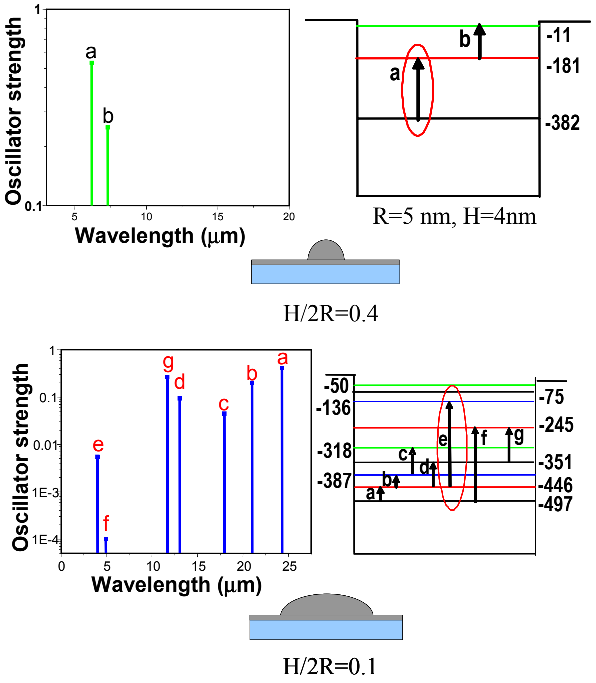


Fig. 11.12. Calculation of energy levels and oscillator strengths for InGaAs dots in InGaP barriers. The dots have different aspect ratios and fixed height.

11.3.5. Quantum dot shape

Along with the size, the shape of the QD is important in determining the energy level structure. "Ideal" QDs are spherical or at least symmetrically shaped, but often this is not the case for SK grown dots, which tend to have pyramidal or lens-like shapes that are flattened in the growth plane (i.e. wider than they are tall). If large enough, this asymmetry can result in a polarization dependent absorption because the confinement will become more QWIP-like. The shape of the QD is not as easily controlled as the QD size.

11.3.6. Quantum dot density

The QD density in SK growth is determined by the QD nucleation process and subsequent ripening. During SK growth, the adatom mobility on the growth surface strongly determines how close together the QDs will nucleate. For example, most growth experiments have shown that decreasing the growth temperature can increase the quantum dot density. Since the adatom mobility is reduced at lower temperature, a given adatom cannot travel as far to find a favorable nucleation center so on average the nucleation centers will be closer together. After the nucleation finishes, additional deposition of material will only grow the existing dots and not nucleate new ones. After nucleation and dot growth and without further deposition of dot material or a cap layer, the dot array will either tend to stabilize or undergo ripening, depending on the growth conditions. Under stable conditions the dots will reach an equilibrium size and spacing. Under conditions that favor ripening the dots will migrate and merge with one another reducing the density and creating dots that are either too large for 3D confinement effects or relax incoherently via defects.

For detector applications a high density of quantum dots above 5×10^{10} dots/cm^2 is desired. For SK growth of In(Ga)As dots on GaAs substrate, densities of up to 10^{11} dots/cm^2 have been demonstrated though densities in the low or mid 10^{10} range are more typical.

The QD density effects both the absorption and transport characteristics of a QD detector. Phillips formulated the absorption coefficient α in a QDIP as follows

Eq. (11.5) $$\alpha(E) = A\frac{n_1}{D}\frac{\sigma_{QD}}{\sigma_{ens}}\exp\left(-\frac{(E-E_G)^2}{\sigma_{ens}^2}\right)[cm^{-1}]$$

where E is the energy, A is the maximum theoretical absorption coefficient, n_1 is the areal density of electrons in the QD ground state, D is the QD density, E_G is the ground to excited state transition energy (for a two level QD system). σ_{QD} and σ_{ens} are the standard deviations in the Gaussian lineshape for intersubband absorption in a single quantum dot and for the distribution in energies for the QD ensemble, respectively. The n_1/D term deals with the reduction in absorption due to lack of electrons in the ground state and the σ_{QD}/σ_{ens} term deals with the reduction in absorption due to dot-to-dot non-uniformity, which will be discussed later. We have stated that a high density of QDs is desired for good performance, however, in Eq. (11.5) the density is in the denominator. This is because density cannot be considered independently of the electron occupation in the dots, which comes into the equation via the n_1 term.

The density also relates to the inter-dot spacing, which effects the amount of lateral interaction between dots in the same layer. Ryzhii's QDIP model predicts that the best QDIP performance occurs for QD arrays that have dots that have a small coupling giving rise to a miniband. [Ryzhii et al. 1996][Ryzhii et al. 2001] This assures that the charge distribution and potential in the plane will be a uniform one. For low dot densities, the QDs behave as localized changes in the potential with the spaces in between acting like punctures where current can flow easily.

11.3.7. Quantum dot uniformity

In a QDIP device the uniformity of the quantum dots has a very strong correlation with device performance. The non-uniformity of the dots can occur in any of the material parameters such as strain, composition, shape, or size, but size is usually considered the most dominant and is also the most easily quantifiable via methods like atomic force microscopy (AFM) or transmission electron microscopy (TEM). One source of size non-uniformity is the presence of incoherent, relaxed dots in an array. These are usually of much lower density than the coherent dots but are significant due to their very large size and potential to create defects in subsequent layers. These defect dots can be almost completely avoided under good growth conditions. The main source of non-uniformity is the Gaussian distribution of the size of the coherent dots. There are two possible "dimensions" of non-uniformity to consider, we first discuss the layer to layer non-uniformity, then we look at the non-uniformity within a single layer.

In a real QDIP, in order to get a high enough volume density of dots it is necessary to stack QD layers on top of one another in the same way QWIPs consist of stacks of QW. The layer grown on top of the QDs to cover them is called the cap layer. Fortunately for device applications, the strain profile of the SK QDs results in preferential growth such that the cap layer returns to a flat surface once the QDs have been completely covered, which facilitates the stacking of more dot layers on top. This flattening process is illustrated in Fig. 11.13.

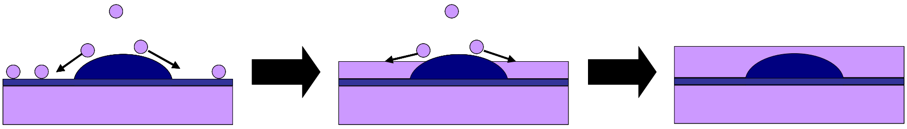


Fig. 11.13. Schematic illustrating the mechanism for flat cap layer growth.

For the stacking of dots the layer-to-layer spacing is important because below a certain thickness the strain field of the buried dot layer will cause the QDs in subsequent layers to vertically align with the dots below. For general detector applications this is not desired because aligned dots tend to increase in size with stack number, have lower spatial coverage due to non-random positioning, and may increase the probability of dot-to-dot tunneling increasing the dark current. For thick enough spacings and if there are neither excess strain nor defects, each dot layers grows more or less independent of one another with each layer having similar in-layer dot uniformity

In the literature, the standard deviation in size for SK grown QDs is around 10%. Optical, electrical, and structural measurements on QD arrays suggests that the variation is a Gaussian distribution. One possible cause for this non-uniformity is asynchronous nucleation of islands due to local variations in the thickness of the QD layer. This amount of non-uniformity can significantly affect the detector performance. Recall from Eq. (11.5) that the relevant measure of non-uniformity is the ratio of the linewidth of a single dot to that of the entire array σ_{QD}/σ_{ens}. Increased non-uniformity in the dot array will broaden the spectral response resulting in a lower peak responsivity (although the integrated responsivity will be similar to a more uniform array). Values of 0.01 for σ_{QD}/σ_{ens} are typical for current fabrication technology, and according to Phillips's modeling results, cause an order of magnitude decrease in detectivity from the ideal case of $\sigma_{QD}/\sigma_{ens} = 1$ (see Fig. 11.14).

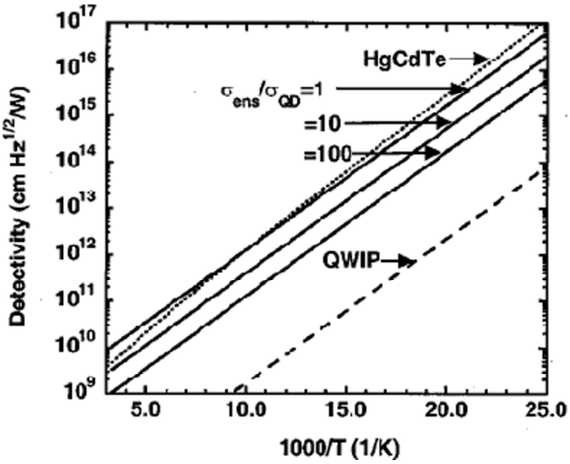


Fig. 11.14. Detectivity for HgCdTe, QWIP, and QDIP detectors as a function of temperature, with the solid lines showing the effect of non-uniformity (σ_{ens}/σ_{QD}) on the QDIP detectors. [Reproduced with permission from Journal of Applied Physics Vol.91, No. 7, J. Phillips, "Evaluation of the fundamental properties of quantum dot infrared detectors," fig. 5, p. 4594, Copyright 2002, American Institute of Physics (AIP).]

11.3.8. Conclusion and future directions for dot fabrication

Despite the shortcomings of SK growth discussed in the sections above, it is still the most successful QD fabrication technology for QDIPs because it can provide defect-free dots. In the next section, we review some of the device results that have been achieved using the technique. The main drawback of self-assembly is its inherently "random" nature that makes it difficult to control the dot characteristics and uniformity. Developing a better understanding of the mechanisms of SK and ways to control it will be critical to achieving the full potential of QDIP performance. One technique being explored to better control self-assembly is growth on pre-patterned substrates. The technique is a hybrid of bottom-up and top-down methods since the pre-patterning will rely on top-down lithographic methods but actual dot growth will still be bottom-up self assembly. In this technique SK QDs are growth either on a patterned surface or on a surface regrown over a patterned surface. The strain fields of these patterned surfaces can potentially be used to direct the growth of the QDs.

11.4. Review of actual QDIP performance

11.4.1. Introduction

In actual demonstrations of QD-based detectors, the performance still falls far short of the theoretical predictions and the maturity level of the technology still lags behind that of the more established QWIP and HgCdTe detectors. This is due largely to the various SK growth limitations discussed in the QD growth section of this chapter. In this section we highlight some of the more recent achievements in QDIP technology in two important areas that are indicative of the maturity of QDIP technology: high operating temperature and FPA imaging.

11.4.2. High operating temperature

The main benefit to be gained from QDIP is high operating temperature. Thus far, QD-based detectors have shown the capability to operate at high temperature including up to room temperature, but the performance is still significantly reduced compared to low operating temperatures.

There have been 4 major demonstrations of high operating temperature (>150 K) QDIPs. The first two devices come from collaborations between the University of Michigan and Georgia State University.[Bhattacharya et al. 2005][Chakrabarti et al. 2005] One detector is a very thick 70 layer InAs QD / GaAs barrier device with an AlGaAs current blocking layer at one end.

This device showed a MWIR response with a peak detectivity in the high 10^9 cm·Hz$^{1/2}$/W at 200 K operating temperature. The device structure and performance are shown in Fig. 11.15.

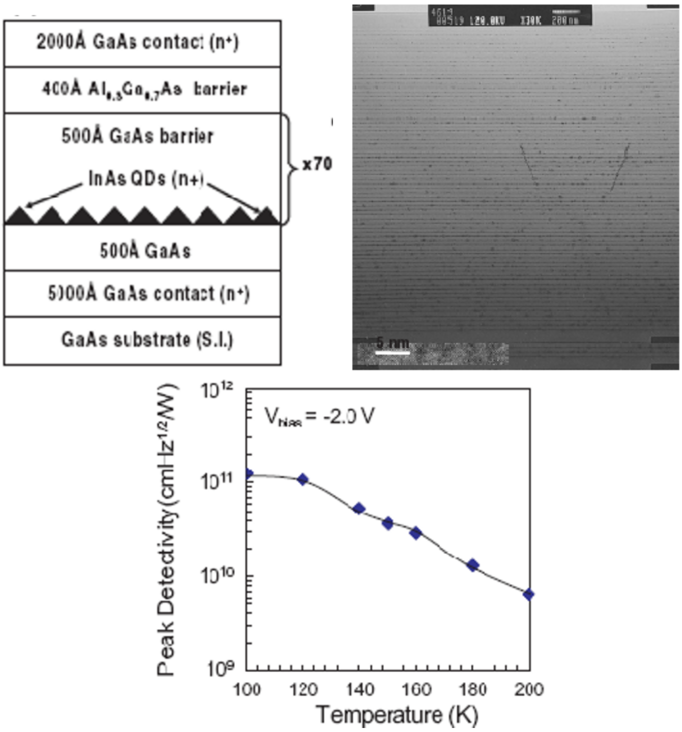


Fig. 11.15. Structure and device performance of the 70 layer QDIP. [Adapted with permission from Journal of Physics D: Applied Physics Vol.38, S. Chakrabarti, A.D. Stiff-Roberts, X.H. Su1, P. Bhattacharya1,G. Ariyawansa, and A.G.U. Perera, "High-performance mid-infrared quantum dot infrared photodetectors," fig. 1 & fig. 3c, p. 2136 & 2139, Copyright 2005, IOP Publishing Ltd.]

The other device from these groups was a more novel design featuring a resonant tunneling dark current filter. This device used InAs QDs in GaAs barriers with AlGaAs layers used as additional barriers and for the resonant tunneling structure. The device showed two detection peaks, one at 6 µm and another at 17 µm. The D* for the 17 µm device was in the low 10^7 cm·Hz$^{1/2}$/W for temperatures from 240 K up to 300 K. The structure and performance are summarized in Fig. 11.16.

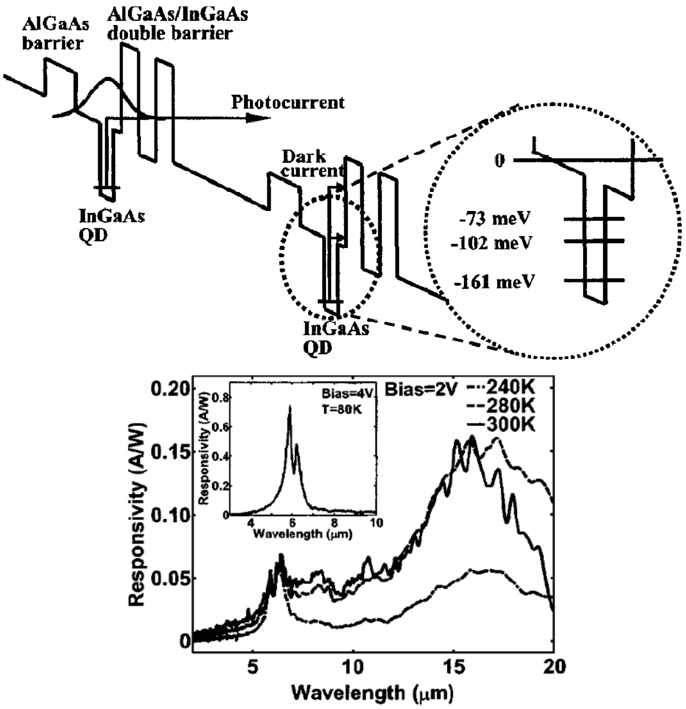


Fig. 11.16. The structure design and photocurrent spectrum of the resonant tunneling barrier QDIP. [Adapted with permission from Applied Physics Letters Vol.86, P. Bhattacharya, X.H. Su, S. Chakrabarti, G. Ariyawansa and A.G.U. Perera, "Characteristics of a tunneling quantum-dot infrared photodetector operating at room temperature," fig. 1 & fig. 3, p. 191106-2, Copyright 2005, American Institute of Physics (AIP).]

It is interesting to note that even though QDs are supposed to inherently provide low dark currents, for both these detectors special dark current reducing structures were used to obtain the high temperature performance.

The other two high operating temperature devices are variants of the dot-in-a-well (DWELL) design. One is from Northwestern University and the other from the University of Massachusetts Lowell and Raytheon Systems [Lim et al. 2007][Lu et al. 2007]. In the Northwestern device InAs QDs embedded in a InGaAs/InP multiple quantum well structure are utilized for MWIR detection at 4 μm. Detection was shown up to room temperature and a D* of 6.7 × 10^7 cm·$Hz^{1/2}$/W at room temperature was reported.

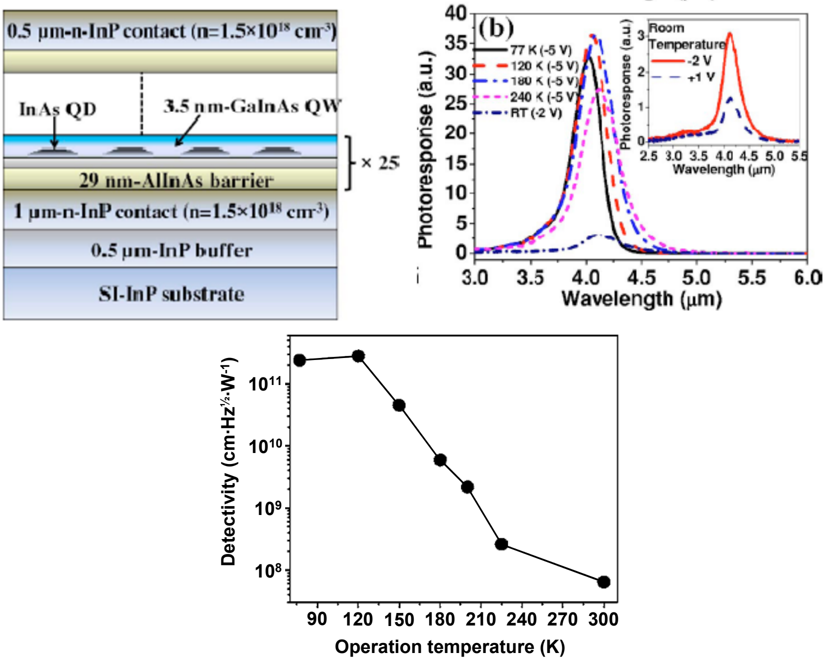


Fig. 11.17. Northwestern University's DWELL structure operating at up to room temperature. [Adapted with permission from Applied Physics Letters Vol.90, H. Lim, S. Tsao, W. Zhang, and M. Razeghi, "High-performance InAs quantum-dot infrared photodetectors grown on InP substrate operating at room temperature," fig. 1,2, & 4, p. 131112-2, Copyright 2007, American Institute of Physics (AIP).]

The University of Massachusetts Lowell and Raytheon Systems device utilized an InAs QD, in an InGaAs/GaAs multiple quantum well structure. The device showed long wavelength performance with peak detection at 9.9 μm. At 190 K the D* was 1.6×10^8 cm·Hz$^{1/2}$/W. The device performance is summarized in Fig. 11.18. This device interestingly showed a reduction in dark current by introducing the quantum dots and without using a current blocking layer.

It is interesting to note that both of these devices are DWELL structures where the QDs are capped with the QW and the barrier is a third material with larger band offset. DWELL structures are popular because they combine some of the benefits of both QWIPs and QDIPs, namely the easy detection wavelength control of QWIPs by adjusting the QW thickness and the normal incidence sensitivity of QDIPs.

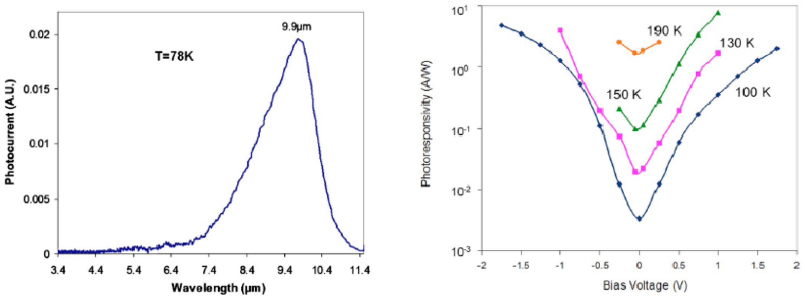


Fig. 11.18. Photoresponse and bias and temperature dependent response of the University of Massachusetts Lowell and Raytheon Systems device structure. [Adapted with permission from Applied Physics Letters Vol.91, X. Lua, J. Vaillancourt, and M.J. Meisner "Temperature-dependent photoresponsivity and high-temperature (190 K) operation of a quantum dot infrared photodetector," fig. 1 & fig. 4, p. 051115-1 & 051115-3, Copyright 2007, American Institute of Physics (AIP).]

11.4.3. FPA imaging

QDIPs have only recently begun to show high performance in FPAs in some cases nearing (but not quite reaching) performance levels comparable to HgCdTe and QWIP FPAs. There are 4 notable QD-based FPA imaging demonstrations. These FPAs either had relatively low NEΔTs for QD-based detectors, 2-color operation, and/or high temperature operation.

Chronologically the first comes from Krishna et al. (2005), which was the first 2-color QD-based FPA. This FPA had bias tunable MWIR and LWIR response and was based on a InAs/InGaAs/GaAs DWELL design. The array was estimated to have an NEΔT < 100 mK and operated at 80 K. Some sample imaging is shown in Fig. 11.19. The two soldering iron images are taken through bandpass filters that demonstrate the multicolor capability of the array. The bottom image shows an infrared image of a person taken with an array of similar design.

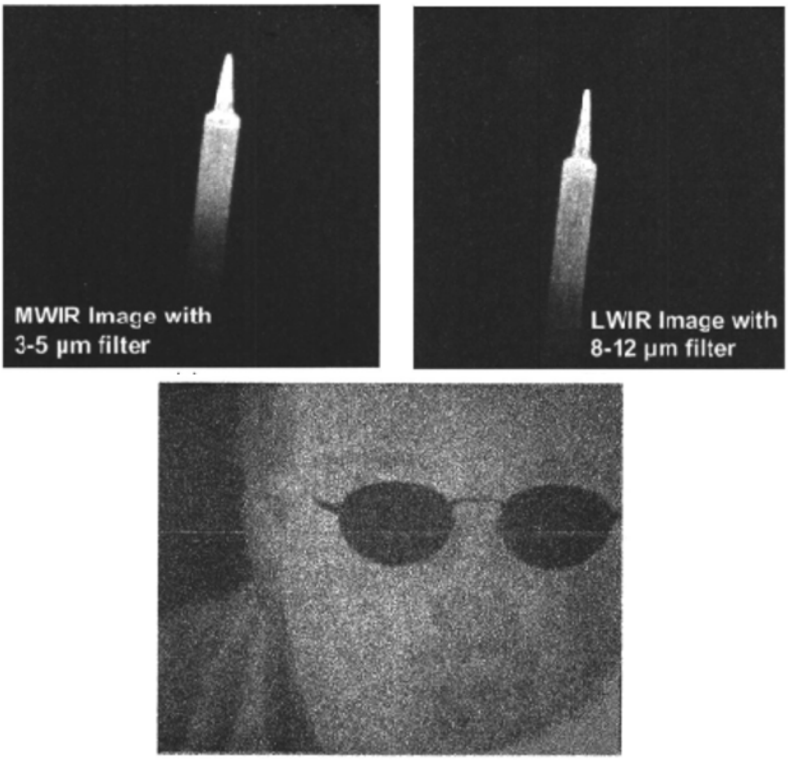


Fig. 11.19. Top: Two images of soldering irons from the 2-color DWELL array by Krishna et al. *The left image is taken through a MWIR filter and the right image is taken through a LWIR filter (both are from the same camera). These images demonstrate the multicolor capability of the array. Bottom: An IR image of a person taken with an array of similar design. [Adapted with permission from Applied Physics Letters Vol.86, S. Krishna, D. Forman, S. Annamalai, P. Do, P. Varangis, T. Tumolillo, A. Gray, J. Zilko, K. Sun, M. Liu J. Campbell, and D. Carothers, "Demonstration of a 320Ã256 two-color focal plane array using InAs/InGaAs quantum dots in well detectors," fig. 3 & fig. 4, p. 193501-2 & 193501-3, Copyright 2005, American Institute of Physics (AIP).]*

Next was a FPA demonstration from Gunapala et al. (Fig. 11.20) [Gunapala et al. 2006]. This was a LWIR DWELL using In(Ga)As QD in GaAs well with AlGaAs barriers. This array is notable because it is the largest format QD-based array demonstrated thus far at 640 × 512. Also the NEΔT was a very low 40 mK when operated at 60 K.

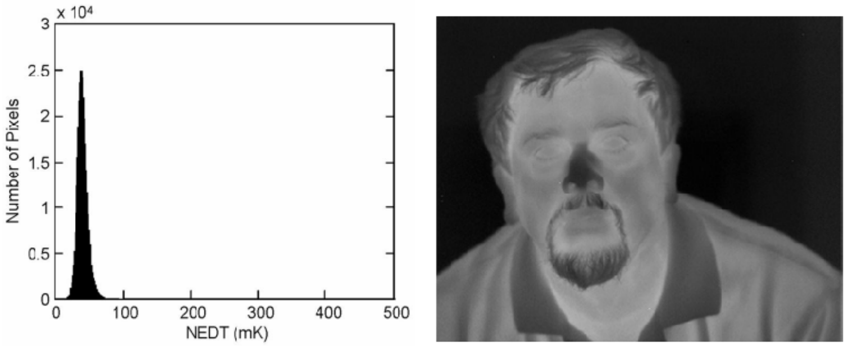


Fig. 11.20. NEΔT histogram and sample imaging from the FPA by Gunapala et al. *[Adapted with permission from Proceedings of the SPIE Vol.6361, S.D. Gunapala, S.V. Bandara, C.J. Hill, D.Z. Ting, J.K. Liu, S.B. Rafol, E.R. Blazejewski, J.M. Mumolo, S.A. Keo, S. Krishna, Y.C. Chang, and C.A. Shott,"640 × 512 pixels Long-Wavelength Infrared (LWIR) Quantum Dot Infrared Photodetector (QDIP) imaging focal plane array," fig. 8 & fig. 10, p. 636116-6 & 193501-7, Copyright 2006, SPIE.]*

The next array to be published (Fig. 11.21 and Fig. 11.22) was from Tsao et al. [2007]. This FPA utilized the structure discussed earlier from Northwestern University. [Lim et al. 2007] The notable accomplishment here was a high maximum operating temperature of 200 K for the FPA due to the low dark current. The reported NEΔT was 344 mK at 130 K.

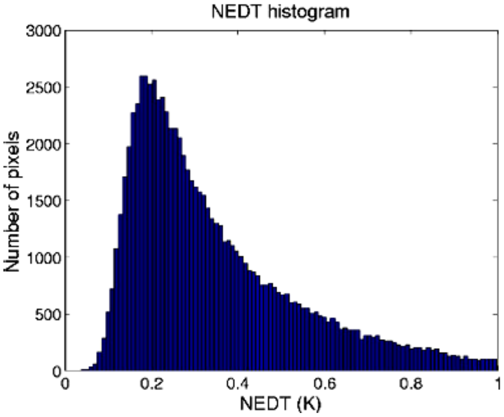


Fig. 11.21. NEΔT histogram from the FPA by Tsao et al[Adapted with permission from IEEE Sensors Journal Vol.8, No. 6, S. Tsao, H. Lim, H. Seo, W. Zhang, and M. Razeghi," InP-Based Quantum-Dot Infrared Photodetectors With High Quantum Efficiency and High-Temperature Imaging," fig. 8, p. 940, Copyright 2008, IEEE.]

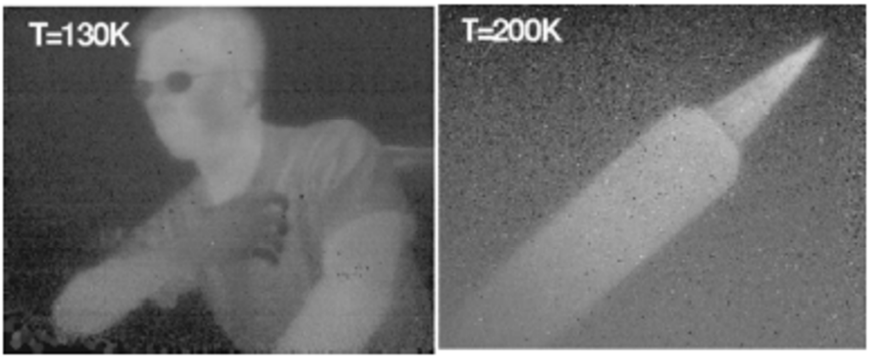


Fig. 11.22. Sample imaging from the FPA by Tsao et al[Adapted with permission from IEEE Sensors Journal Vol.8, No. 6, S. Tsao, H. Lim, H. Seo, W. Zhang, and M. Razeghi," InP-Based Quantum-Dot Infrared Photodetectors With High Quantum Efficiency and High-Temperature Imaging," fig. 7, p. 939, Copyright 2008, IEEE.]

Finally, the most recent QD-based FPA report comes from Varley et al. [2007] (Fig. 11.23 and Fig. 11.24). This FPA is also based on an InAs/InGaAs/GaAs DWELL structure. This is a 2 color, bias tunable FPA showing response in the MWIR and LWIR with NEΔT of 55 mK and 70 mK, respectively at 77 K.

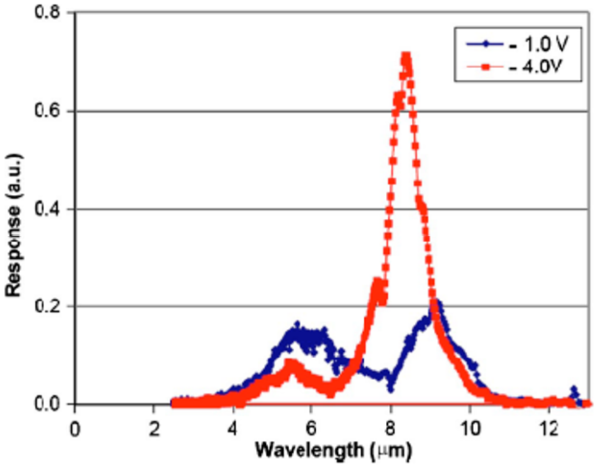


Fig. 11.23. Photoresponse spectrum from the 2 color DWELL FPA by Varley et al. *[Adapted with permission from Applied Physics Letters Vol.91, E. Varley, M. Lenz, S.J. Lee, J.S. Brown, D.A. Ramirez, A. Stintz, S. Krishna, A. Reisinger, and M. Sundaram, "Single bump, two-color quantum dot camera," fig. 1, p. 081120-2, Copyright 2007, American Institute of Physics (AIP).]*

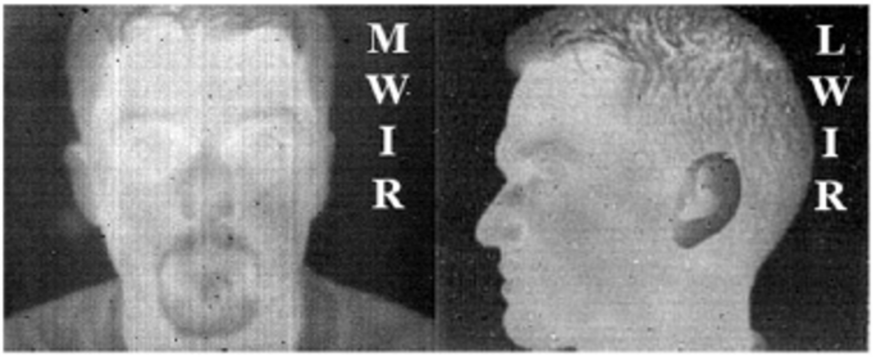


Fig. 11.24. Sample imaging from the 2 color DWELL FPA by Varley et al. *[Adapted with permission from Applied Physics Letters Vol.91, E. Varley, M. Lenz, S.J. Lee, J.S. Brown, D.A. Ramirez, A. Stintz, S. Krishna, A. Reisinger, and M. Sundaram, "Single bump, two-color quantum dot camera," fig. 4, p. 081120-3, Copyright 2007, American Institute of Physics (AIP).]*

11.4.4. Summary

From the high temperature operation and varied imaging demonstrations discussed in the previous section, it is clear that QD-based detectors are being seriously studied as candidates for the next generation of infrared detectors and that the potential is there. However, the leading performance devices all either contain structures that either block dark current or utilize dots in wells. Pure QDIP device results have been demonstrated but they are generally inferior to the results just discussed. This means that at the present level of the technology we are still compensating for imperfect QD characteristics with other clever tricks. The theoretical potential for these alternative designs has not been explored as thoroughly as that for pure QDIPs. And in any case the demonstrated performance for these alternative devices is still far short of the predictions, leaving room for growth and improvement.

11.5. Summary

In this chapter we began with the basic operating principles of QWIPs and QDIPs as a background for the discussion on the theoretical merits of QDIPs. This was followed by a look at the growth technology for creating QDs and how this technology still faces challenges in achieving QDs that can provide the full potential of QD detectors. Finally, we finished with a brief review of the state-of-the-art in QD-detectors and noted that none of the devices are pure QDIPs because it is still necessary to design around the shortcomings of current QDIP devices.

Since their initial proposal however in the mid-1990s, QDIPs have made a great deal of progress to get to where they are now with active research and development of the technology, and with continuing advances in QD growth and fabrication technology, the full potential of QDIP devices may be achievable in the near future.

References

Bhattacharya, P., Su, X.H. and Chakrabarti, S., "Characteristics of a tunneling quantum-dot infrared photodetector operating at room temperature," *Applied Physics Letters* 86, pp. 191106-1, 2005.

Bimberg, D., Grundmann, M. and Ledenstov, N.N., *Quantum dot heterostructures*, John Wiley & Sons Ltd., Chichester, 1999.

Chakrabarti, S., Stiff-Roberts, A.D., Su, X.H., Bhattacharya, P., Ariyawansa, G. and Perera, A.U., "High-performance mid-infrared quantum dot infrared photodetectors," *Journal of Physics D: Applied Physics* 38, pp. 2135-2141, 2005.

Goldstein, L., Glas, F., Marzin, J.Y., Charasse, M.N. and Roux, G.L., "Growth by molecular beam epitaxy and characterization of InAs/GaAs strained-layer superlattices," *Applied Physics Letters* 47, pp. 1099-1101, 1985.

Gunapala, S.D., Bandara, S.V., Hill, C.J., Ting, D.Z., Liu, J.K., Rafol, S.B., Blazejewski, E.R., Mumolo, J.M., Keo, S.A., Krishna, S., Chang, Y.C. and Shott, C.A., "640×512 pixels long-wavelength infrared (LWIR) quantum dot infrared photodetector (QDIP) imaging focal plane array," *Proceedings of SPIE* 6361, pp. 636116, 2006.

Krishna, S., Forman, D., Annamalai, S., Dowd, P., Varangis, P., Tumolillo, T., Jr, Gray, A., Zilko, J., Sun, K., Liu, M., Campbell, J. and Carothers, D., "Demonstration of a 320×256 two-color focal plane array using InAs/InGaAs quantum dots in well detectors," *Applied Physics Letters* 86, pp.193501-1, 2005.

Ledentsov, N.N., Shchukin, V.A., Bimberg, D., Ustinov, V.M., Cherkashin, N.A., Musikhin, Y.G., Volovik, B.V., Cirlin, G.E. and Alferov, Z.I., "Reversibility of the island shape, volume and density in Stranski-Krastanow growth," *Semiconductor Science and Technology* 16, pp. 502-506, 2001.

Lim, H., Tsao, S., Zhang, W. and Razeghi, M., "High-performance InAs quantum-dot infrared photodetectors grown on InP substrate operating at room temperature," *Applied Physics Letters* 90, pp. 131112-1, 2007.

Lu, X., Vaillancourt, J. and Meisner, M.J., "Temperature-independent photoresponsivity and high-temperature (190 K) operation of a quantum dot infrared photodetector," *Applied Physics Letters* 91, pp. 051115-1, 2007.

Martyniuk, P., Krishna, S. and Rogalski, A., "Assessment of quantum dot infrared photodetectors for high temperature operation," *Journal of Applied Physics* 104, pp. 034314-1, 2008.

Martyniuk, P. and Rogalski, A., "Quantum-dot infrared photodetectors: Status and outlook," *Progress in Quantum Electronics* 32, pp. 89-120, 2008.

Phillips, J., "Evaluation of the fundamental properties of quantum dot infrared detectors," *Journal of Applied Physics* 91, 4590-4594, 2002.

Ryzhii, V., "The theory of quantum-dot infrared phototransistors," *Semiconductor Science and Technology* 11, pp. 759-765, 1996.

Ryzhii, V., Irina, K., Maxim, R., Victor, I.P., Vladimir, V.M. and Magnus, W., "Why QDIPs are still inferior to QWIPs: theoretical analysis," *Proceedings of SPIE* 4288, pp. 396-403, 2001.

Tsao, S., Lim, H., Seo, H., Zhang, W. and Razeghi, M., "InP-Based Quantum-Dot Infrared Photodetectors With High Quantum Efficiency and High-Temperature Imaging," *IEEE Sensors Journal* 8, pp. 936-941, 2008.

Varley, E., Lenz, M., Lee, S.J., Brown, J.S., Ramirez, D.A., Stintz, A., Krishna, S., Reisinger, A. and Sundaram, M., "Single bump, two-color quantum dot camera," *Applied Physics Letters* 91, pp. 081120-1, 2007.

Further reading

Bimberg, D., Grundmann, M. and Ledenstov, N.N., *Quantum dot heterostructures*, John Wiley & Sons Ltd., Chichester, 1999.

Madhukar, A., Kim, E-T., Chen, Z., Campbell, J. and Ye, Z., "Quantum Dot Infrared Photodetectors," in *Semiconductor Nanostructures for Optoelectronic Applications,* ed. T.D. Steiner, pp. 45-112, Artech House, Boston, 2004.

Martyniuk, P. and Rogalski, A., "Quantum-dot infrared photodetectors: Status and outlook," *Progress in Quantum Electronics* 32, pp. 89-120, 2008.

Razeghi, M., Zhang, W., Lim, H., and Tsao, S., "Quantum Dot Infrared Photodetectors by Metal-Organic Chemical Vapor Deposition," in *Handbook of Self Assembled Semiconductor Nanostructures for Novel Devices in Photonics and Electronics*, ed. M. Henini, pp. 620-658, Elsevier Limited, Amsterdam, 2008.

Problems

1. Calculate the minimum dot size for electrons and holes for the follow QD material systems InAs/GaAs, InAs/InP, InAs/InGaAs lattice matched to GaAs.

2. Assume quantum dots are hemispherical and always arrange in square arrays.

3. What is the maximum density for QDs 15 nm in diameter?

4. What is the inter-dot spacing (center to center) for an array with QD density of 10^{10} dots/cm^2 and size of 30 nm?

5. What is the fill factor for the arrays in parts a and b?

6. Given a QD non-uniformity with standard deviation of 10% for a mean QD size of 15 nm and QD density 1×10^{11} dots/cm^2, what kind of pixel-to-pixel non-uniformity would you expect to see in the absorption strength for each 25 μm × 25 μm pixels? (Hint: Consider the number of dots per pixel).

7. In which material system might you expect QD to form more easily, InAs/InP or InAs/GaAs? Justify your answer.

8. Let us roughly estimate the effect of broadening on the peak absorption. A series of QDIP detectors have the same integrated blackbody responsivity, the same peak detection wavelength of 9 μm, and the normalized spectral responses have Gaussian-shaped detection spectra but with different full width at half maximums due to dot non-uniformity. The 3 detectors have full width half maximums of 10%, 20%, and 30%. Assuming the same integrated absorption for all three detectors, calculate how does the relative variation of peak absorption intensity for the three detectors.

12. Single-Photon Avalanche Photodiodes

12.1. Introduction
12.2. Avalanche photodetectors, linear mode
  12.2.1. Device fabrication
  12.2.2. Linear-mode operation
  12.2.3. Excess noise
12.3. Examples of APD structures
  12.3.1. Reach-through avalanche photodiodes
  12.3.2. Separate absorption charge multiplication (SACM) APD
12.4. Geiger Mode Operation
  12.4.1. Basic theory
  12.4.2. Passive avalanche quenching
  12.4.3. Active avalanche quenching
  12.4.4. Gated detection
  12.4.5. Device limitations
  12.4.6. After-pulsing
12.5. Examples of single-photon avalanche photodiodes
  12.5.1. Silicon single-photon avalanche diodes
  12.5.2. InGaAs/InP single-photon avalanche diodes
  12.5.3. GaN single-photon avalanche diodes
12.6. Summary

12.1. Introduction

The detection of single photons has attracted the attention of scientists for many years. Applications such as Raman spectroscopy, fluorescence spectroscopy, or quantum key distribution require the use of devices with such level of sensitivity. Thanks to their high internal gain, photomultiplier tubes were the first devices to demonstrate single-photon counting capabilities. However, their high volume and required voltages soon encouraged the search for new devices.

The impact ionization phenomenon was intensively studied in Silicon, Germanium, GaP and GaAs during the 1950's and 1960's, and several

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_12,
© Springer Science+Business Media, LLC 2010

theories accounted for the dependence of the ionization rate upon the electric field in *p-n* diodes near the breakdown voltage [Wolf 1954][Shockley 1961][Baraff 1962]. Practical advantages of these devices were predicted in cases in which the sensitivity of the detection system was limited by the thermal or amplifier noise and not by the shot noise of the detector. In those cases, the use of these high-gain devices would help to improve the signal-to-noise ratio considerably. In 1966, McIntyre published his famous expressions for the calculation of the multiplication noise under selective injection of electrons or holes in avalanche photodiodes with uniform multiplication regions [McIntyre 1966]. A few years later, Webb and McIntyre proposed the use of these devices in photon-counting mode to compete with the available photomultipliers at that time [Webb 1970].

Unfortunately, material issues at that time made the devices often show hot points and microplasmas in avalanche regime [Chynoweth 1956], giving rise to instabilities and lack of uniformity of the optical response. In spite of these drawbacks, reach-through avalanche photodiodes (APDs) became commercially available in the mid-1970s, although they would not be a serious alternative to the photomultiplier tubes untill the late 1980s. During the 1990s, the devices started to form part of complex systems for the measurement of low-power emissions. Hence, its use in photon-correlation spectroscopy, astronomy and high-energy physics became more and more popular [Nightingale et al. 1991] [Daudet et al. 1993][Gullikson 1995]. On the other hand, the novel development of quantum cryptography fostered their use in quantum key distribution experiments [Bennet et al. 1992].

Nowadays, material progress has led to the development of improved avalanche photodiodes with single-photon detection capabilities in traditional semiconductors, such as Si or InGaAs, as well as in novel wide-bandgap technologies. Thus, integrated photon counting systems based on Si single-photon avalanche diodes (SPADs) are today commercially available for a wide spectral range from 350 nm to 900 nm; commercial InGaAs/InP avalanche photodiodes have been successfully tested as single-photon detectors at telecommunication wavelengths [Hiskett et al. 2001]; and in the ultraviolet range, SiC and GaN avalanche photodiodes have demonstrated single-photon detection capabilities [Bai et al. 2007][Pau et al. 2007].

In this chapter, we start reviewing the basic properties of avalanche photodiodes. In the second part, we focus on the main characteristics and issues of Geiger mode operation for photon counting purposes. Towards the end of the chapter, we provide some examples of the state-of-the-art of single-photon avalanche diodes in Si, InGaAs, and GaN.

12.2. Avalanche photodetectors, linear mode

12.2.1. Device fabrication.

Most of the current APD structures designed for Geiger-mode operation are comprised of layers grown by epitaxial techniques such as MBE or MOCVD (Chapter 1, Sections 1.5 & 1.6). Moreover, the high-quality material required to avoid multiplication non-uniformities makes necessary the use of lattice-matched substrates. Hence, for Si photodiodes, the epilayers are grown homoepitaxially on Si substrates, whereas for $In_{0.53}Ga_{0.47}As$, InP substrates are used. The lack of available lattice-matched substrates for other materials like III-nitrides has major consequences such as low available device areas, formation of microplasmas or early breakdown.

The most basic structure of an avalanche photodiode is a *p-i-n* junction. Therefore, many of the processing steps of these devices are common to those followed to fabricate low-voltage photodiodes: mesa etching, surface passivation, metal contact deposition, dopant diffusion and implantation. More details on these steps can be found in Chapters 2 and 3.

Most of the single avalanche photodiodes commercially available rely on planar structures that do not require mesa etching, which considerably reduces the edge effects and eliminates surface damage and device instabilities introduced by the etching. Guard rings are usually incorporated to the APD design with the aim of preventing early breakdown at the periphery of the diode (see Fig. 12.1). Another method to reduce the edge effects in planar InGaAs/InP APDs is the use of a double diffusion process [Liu et al. 1992]. The resultant p^+ area has a thicker intrinsic layer underneath the edges reducing the electric field. In single-photon avalanche diodes (SPADs), this has been demonstrated to result in a reduction of the dark count probability.

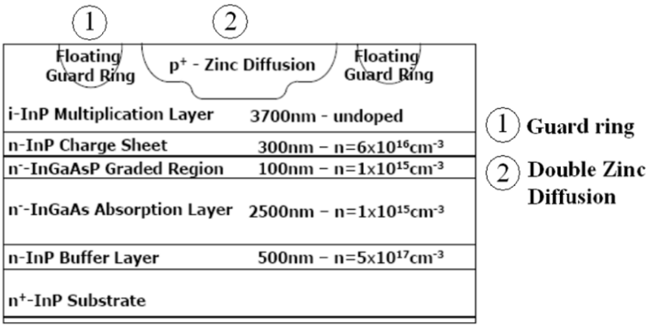

①
②
Floating Guard Ring
p⁺ - Zinc Diffusion
Floating Guard Ring
i-InP Multiplication Layer 3700nm - undoped
n-InP Charge Sheet 300nm – n=6x10¹⁶cm⁻³
n⁻-InGaAsP Graded Region 100nm – n=1x10¹⁵cm⁻³
n⁻-InGaAs Absorption Layer 2500nm – n=1x10¹⁵cm⁻³
n-InP Buffer Layer 500nm – n=5x10¹⁷cm⁻³
n⁺-InP Substrate
① Guard ring
② Double Zinc Diffusion

Fig. 12.1. Schematic cross-section of a planar InGaAs/InP SPAD with floating guard ring, and double diffused junction. [Reproduced from permission of Conference on Lasers and Electro-Optics (CLEO-2006), R.E. Warburton, S. Pellegrini, J.P.R. David, J.S. Ng, A. Krysa, K. Groom, L. Tan, S.D. Cova, G.S. Buller, "Design, fabrication and characterization of InGaAs/InP single-photon avalanche diode detectors", fig. 1, pp.1, Copyright 2006 Optical Society of America (OSA).]

Material doping via diffusion or implantation in wide-bandgap semiconductors is difficult due to the high bond energy, which disables the fabrication of guard rings by these methods. Therefore, mesa etching is mainly used to fabricate APD photodiodes in these materials. In that case, dielectric passivation is crucial to minimize surface effects. Moreover, alternative methods are followed in order to reduce the electric field at the edges like the use of beveled structures. Beveling the sidewalls of the mesa, the depletion region becomes broader at the edges, which makes the electric field in this region lower than the electric field in the bulk, as illustrated in Fig. 12.2.

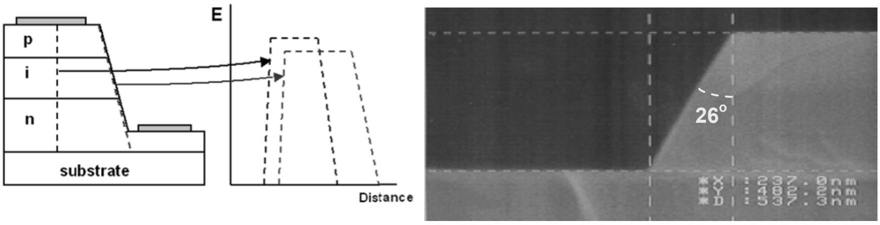

p
i
n
substrate
E
Distance
26°
*X : 237.0nm
*Y : 482.2nm
*D : 537.3nm

Fig. 12.2. Left: Effect of the beveled sidewall of a mesa structure on the electric field profile. Right: Beveled edge of a GaN APD with an angle of 26°.

Another usual step is the inclusion of a recessed window in the structure when the devices are front-illuminated. Its objective is to address issues related to the absorption properties of the material at shorter wavelengths than the bandgap. As the wavelength decreases, the absorption coefficient typically increases. Those photons are absorbed closer to the top semiconductor surface whose energy levels act as recombination centers, hindering the collection of carriers in the depletion region. For high surface recombination rates, the effect on the quantum efficiency at shorter wavelengths is dramatic. Therefore, the use of recessed windows reduces the distance with the depletion region (see Fig. 12.3), allowing better collection efficiency and enhancing the optical response; consequently, the single-photon detection efficiency of SPADs is significantly improved at shorter wavelengths.

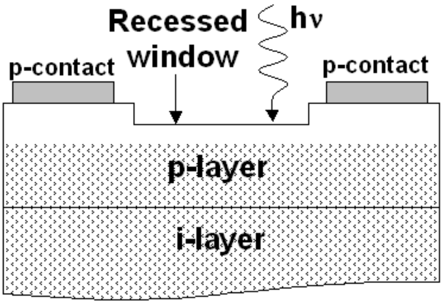


Fig. 12.3. Schematic of the top layers of an APD with recessed window. The dotted area represents the depletion region.

12.2.2. Linear-mode operation.

Material breakdown takes place when the electric field reaches the so-called critical electric field in the semiconductor. This electric field can be determined theoretically and limits the maximum bias voltage that can be applied to a device. In *p-n* diodes, the electric field at the junction increases with reverse voltage. At the breakdown voltage, the electric field strength at the junction equals the critical electric field and there is a strong increase of the current flowing through the device. In an avalanche photodiode, this increase is due to the multiplication of carriers through impact ionization events as they travel across the junction.

It is noticeable that other mechanisms such as Zener tunneling or trap-assisted tunneling could also lead to enhance the current. However, avalanche breakdown present a characteristic behavior as a function of temperature that makes it distinguishable from those cases. Due to the increase of the phonon population, the phonon scattering becomes more likely in the semiconductor as the temperature increases, reducing the probability of impact ionization events in the multiplication region. For that reason, the breakdown voltage at which the avalanche occurs increases with temperature [Minder et al. 2007]. On the other hand, the alternative mechanisms mentioned above present a reduction of the breakdown voltage as the temperature increases.

In *p-i-n* photodiodes, the insertion of an intrinsic region between the *p*- and the *n*-layers makes possible to create a uniform electric field profile in which the carriers can be multiplied at a constant rate. The thicker the intrinsic region, the lower is the electric field strength, as shown in Fig. 12.5. Consequently, the breakdown voltage would increase as the intrinsic layer gets thicker [McClintock et al. 2007].

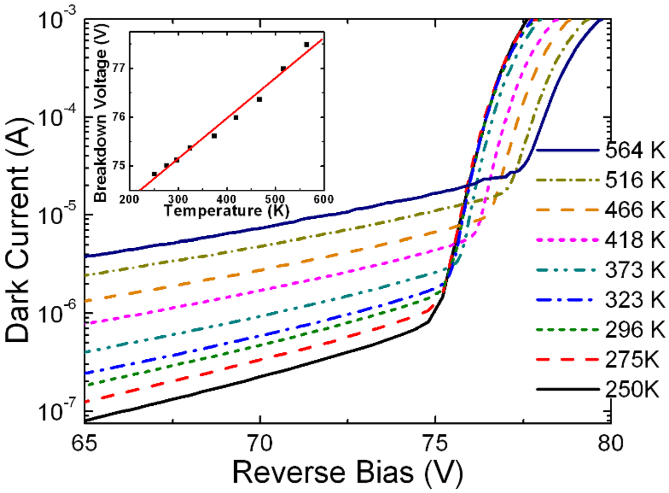


Fig. 12.4. Breakdown voltage increase in GaN APDs as the temperature increases. [Reproduced with permission from Applied Physics Letters Vol. 91, K. Minder, J.L. Pau, R. McClintock, P. Kung, C. Bayram, M. Razeghi, D. Silversmith, "Scaling in Back-Illuminated GaN Avalanche Photodiodes," fig. 5, p. 073513-3, Copyright 2007, American Institute of Physics (AIP).]

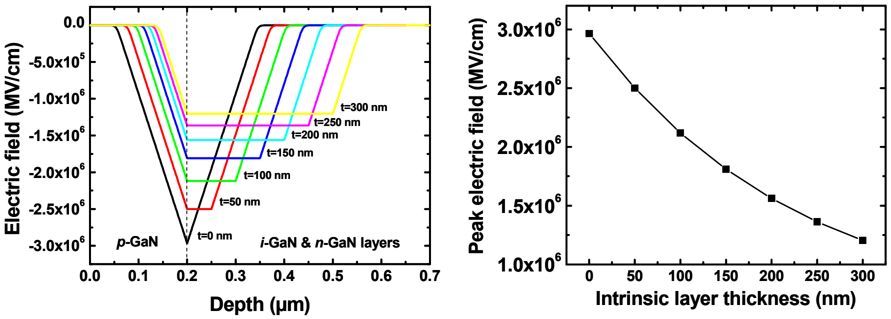


Fig. 12.5. Right: Electric field profile of p-i-n GaN diodes with different intrinsic region thickness (t). Left: Plot of the peak electric field as a function of the intrinsic layer thickness.

The rates at which holes and electrons impact ionize are characteristic of the type of carrier and dependent on the electric field strength. Among the available models to account for the relationship between the ionization rates and the electric field, the most accepted one is the Chynoweth's law. In this model, the ionization rates are expressed as:

Eq. (12.1) $$\alpha,\beta = A(T)\exp\left\{-\left(\frac{B(T)}{E}\right)\right\}$$

where α and β are the ionization coefficients for electrons and holes, respectively, T is the temperature, and E is the electric field. A and B are temperature dependent coefficients that can be determined by fitting the experimental data. For some semiconductors, the ratio $k = \beta/\alpha$ is a constant independent of the electric field; for convenience to interpret the experimental data, this approximation is also used for certain ranges of electric field in other semiconductors.

In avalanche photodiodes, the injection of n photogenerated carriers in the high electric field region of width W results in a total number of m free carriers, where m/n is known as multiplication gain or simply gain. Two cases can be distinguished depending on the type of carrier that initiates the multiplication process. Thus, gain is named M_n when the carrier multiplication is initiated by electrons, and $M_{p,}$ when it is initiated by holes. For electron injection at $x = 0$ and hole injection at $x = W$, both parameters can be calculated from the ionization rates as

Eq. (12.2) $$M_n^{-1} = 1 - \int_0^W \alpha \exp\left[-\int_0^x (\alpha - \beta)\, dx'\right] dx$$

Eq. (12.3) $$M_p^{-1} = 1 - \int_0^W \beta \exp\left[-\int_x^W (\alpha - \beta)\, dx'\right] dx$$

In particular, when $\alpha = \beta$ in a uniform electric field:

Eq. (12.4) $$M_n = M_p = \left[1 - \int_0^W \alpha\, dx\right]^{-1} = (1 - \alpha W)^{-1}$$

In case $\alpha \neq \beta$ but the electric field is uniform enough to consider the ionization coefficients independent of the distance across the depletion region, the multiplication factors can be approximated as

Eq. (12.5) $$M_n = \frac{[1-(\beta/\alpha)]\exp\{\alpha W[1-(\beta/\alpha)]\}}{1-(\beta/\alpha)\exp\{\alpha W[1-(\beta/\alpha)]\}}$$

Eq. (12.6) $$M_p = \frac{[1-(\beta/\alpha)]}{1-(\beta/\alpha)\exp\{\alpha W[1-(\beta/\alpha)]\}}$$

For injection of both types of carriers in the high-field region, the multiplication factor can be averaged as

Eq. (12.7) $$\langle M \rangle = \frac{I_n M_n + I_p M_p}{I_t}$$

where I_n, I_p and I_t are the electron, hole and total current injected. Thus, when the APD is biased at voltages close to the breakdown voltage, the photogenerated current ($I_{m,ph}$) is amplified with a gain equal to $\langle M \rangle$ with respect to the photocurrent at low voltages ($I_{0,ph}$):

Eq. (12.8) $$I_{m,ph} = \langle M \rangle I_{0,ph} = \langle M \rangle \frac{q\lambda\eta}{hc} P_{in}$$

where P_{in} is the incident power. This mode of operation is called linear mode in contrast to the Geiger mode that will be presented in Section 12.4.

A typical procedure to determine the ionization coefficients is to compare the multiplication gains of *p*-*i*-*n* avalanche photodiodes under illumination of the *p*-GaN layer to selectively inject electrons (M_n) and under illumination of the *n*-GaN layer to selectively inject holes (M_n). If the electric field (E) is uniform, the ionization coefficients can be obtained as a function of the multiplication gains from Eq. (12.5) and Eq. (12.6) as

Eq. (12.9) $$\beta(E) = \frac{1}{W}\left(\frac{M_p(V)-1}{M_p(V)-M_n(V)}\right)\ln\left(\frac{M_p(V)}{M_n(V)}\right)$$

Eq. (12.10) $$\alpha(E) = \frac{1}{W}\left(\frac{M_n(V)-1}{M_n(V)-M_p(V)}\right)\ln\left(\frac{M_n(V)}{M_p(V)}\right)$$

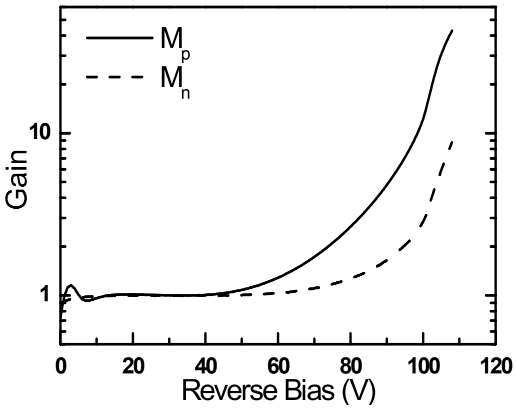


Fig. 12.6. Multiplication gains obtained in GaN p-i-n avalanche photodiodes with selective injection of holes (M_p) and electrons (M_n) in the depletion region. [Reproduced with permission from Applied Physics Letters Vol. 90, R. McClintock, J.L. Pau, K. Minder, C. Bayram, P. Kung, M. Razeghi, "Hole-initiated multiplication in back-illuminated GaN avalanche photodiodes," fig. 2, p. 141112-2, Copyright 2007, American Institute of Physics (AIP).]

12.2.3. Excess noise.

The linear-mode operation of APD has intrinsically associated a characteristic type of noise called excess noise. This noise source has its origin in the statistical nature of the multiplication process. As not all the carriers yield the same number of ionization events n in their travel across the high-field region, the average of the squared multiplication factor $\langle n^2 \rangle$ is not equal to the square of the averaged multiplication factor $\langle n \rangle^2$. Therefore, the so-called Excess Noise Factor (F) is defined as $\langle n^2 \rangle / \langle n \rangle^2$. Following McIntyre's theory, the resultant noise spectral density can be calculated as

Eq. (12.11) $$S = 2qI_0 \langle M \rangle^2 F$$

where I_0 is the total current, and $\langle M \rangle$ is the averaged multiplication gain. In those cases in which $k = \beta/\alpha$ can be approximated by a constant, F is given by:

Eq. (12.12) $$F = M_p \left[1 + \left(\frac{1-k}{k} \right) \left(\frac{M_p - 1}{M_p} \right) \right]^2$$

if the only carriers injected in the depletion region are holes, and by

Eq. (12.13) $F = M_n \left[1 - (1-k) \left(\frac{M_n - 1}{M_n} \right) \right]^2$

if the only injected carriers are electrons.

It is important to notice that when $k > 1$, F is lower if the carriers injected in the depletion region are holes; on the contrary, if $k < 1$, F would be lower if the carriers injected are electrons. In order to minimize the noise in the APD, this must be taken into consideration when new device structures are designed in a particular material system with known α and β coefficients.

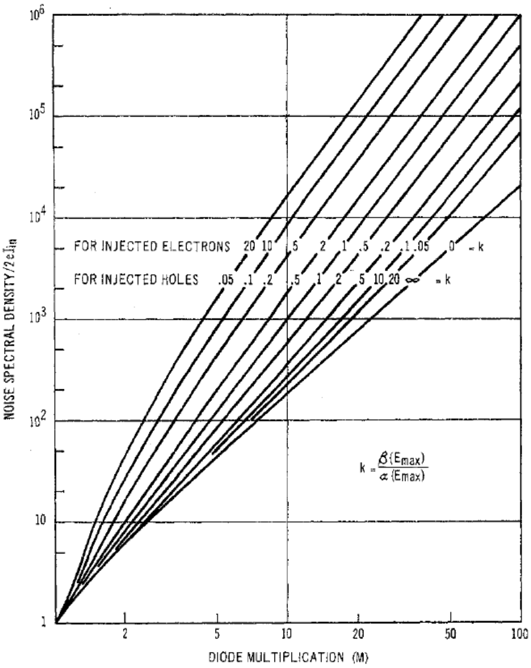


Fig. 12.7. Noise spectral density/2qI_0 vs. multiplication for either injected holes or electrons if β=kα. [Reproduced with permission from IEEE Transactions on Electron Devices Vol. 13, R.J. McIntyre,, "Multiplication Noise in Uniform Avalanche Diodes," fig. 1, p. 168, Copyright 1966, IEEE.]

12.3. Examples of APD structures

Unfortunately, the variety of structures tested during more than fifty years of research in this kind of devices is wide enough to make extremely tedious their revision. Therefore, it is out of the scope of this book to provide details on all of those but to give a couple of examples of structures of supreme importance for their significant role in the development of Si APDs for the visible and near-IR ranges (reach-through avalanche photodiodes) and InGaAs/InP APDs for telecom wavelengths (separate absorption and multiplication APDs).

12.3.1. Reach-through avalanche photodiodes.

These photodiodes rely on n-p-π-p^+ structures. The π and p layers are epitaxially grown on a p^+ substrate that acts as a bottom layer. The n-type region is made by dopant diffusion. The electric field profile of the structure, shown in Fig. 12.8, presents two main regions. The first has low electric field and is known as absorption region because is where most of the photons are absorbed. The second is the high electric field region situated around the junction between the p- and n-type layers. This is known as multiplication region because in it the carriers gain enough energy to produce new electrons-hole pairs by impact ionization. As the illumination takes place in the π- and p-type layers, the carriers injected into the multiplication region are the minority carriers in those layers, i.e. electrons. Therefore, in silicon, this design takes advantage of the highest ionization coefficient for electrons to reduce noise and enhance gain.

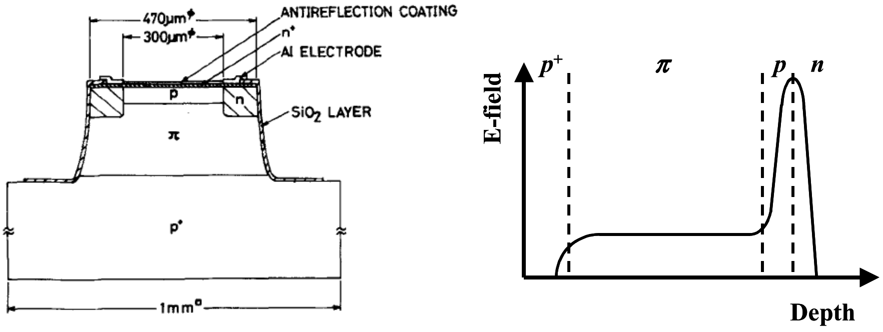


Fig. 12.8. Left: Schematic view of a reach-through Si avalanche photodiode. Right: Electric field profile of a reach-through avalanche photodiode. [Adapted with permission from ", Journal of Applied Physics Vol. 48, M. Maeda, Y. Minai, M. Tanaka, "Photoinduced anomalous oscillations in reach-through avalanche photodiodes," fig. 1, pp.5324, Copyright 1977, American Institute of Physics (AIP).]

12.3.2. Separate absorption charge multiplication (SACM) APD

This structure is used in the fabrication of commercial InGaAs/InP APDs for telecommunication wavelengths (1.3 and 1.55 μm) and relies on the possibility to combine lattice-matched binary, ternary and quaternary alloys for band engineering. The band diagram is shown in Fig. 12.9. Light is absorbed in the InGaAs (absorption region). As the In content needed to lattice-match the InP substrate is 53%, absorption spectrum extends up to $\lambda = 1.65$ μm. The electric field distribution makes possible the injection of photogenerated holes into InP-based multiplication region, whose k value is about 3. To access the multiplication region, the holes must overcome the barriers formed by the valence band offset between InGaAs and InP. To make the transition between both materials smoother, an InGaAsP grading layer is used, which also alleviates hole trapping at the interface. A totally depleted *n*-type InP layer contributes to enhance the electric field in the InP multiplication region. It is usually called "charge layer".

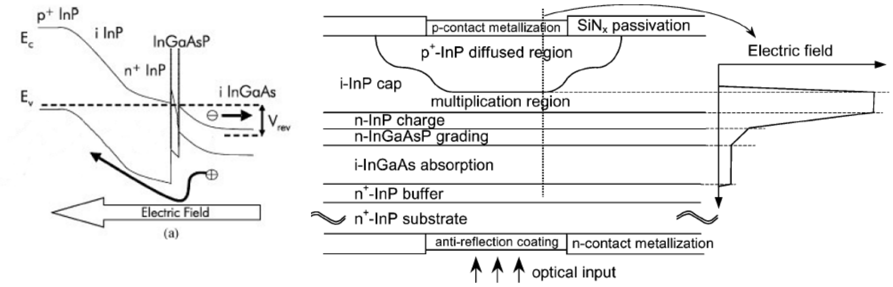

p$^+$ InP
E_c
i InP
InGaAsP
n$^+$ InP
i InGaAs
E_v
V_{rev}
Electric Field
(a)
p-contact metallization
SiN$_x$ passivation
p$^+$-InP diffused region
i-InP cap
multiplication region
n-InP charge
n-InGaAsP grading
i-InGaAs absorption
n$^+$-InP buffer
n$^+$-InP substrate
anti-reflection coating
n-contact metallization
optical input
Electric field

Fig. 12.9. Left: Band diagram of a InGaAs/InP SACM-APD. Right: Schematic cross-section and electric field profile. [Adapted with permission from IEEE Journal of Quantum Electronics Vol. 42, S. Pellegrini, R.E. Warburton, L.J. Tan, J.S. Ng, A.B. Krysa, K. Groom, J.P.R. David, S. Cova, M.J. Robertson, and G.S. Buller, "Design and Performance of an InGaAs–InP Single-Photon Avalanche Diode Detector," fig. 1, pp. 398, Copyright 2006, IEEE. And adapted with permission from IEEE Journal of Selected Topics in Quantum Electronics Vol. 13, No. 4, X. Jiang, M.A. Itzler, E. Ben-Michael, and K. Slomkowski, "InGaAsP–InP Avalanche Photodiodes for Single-Photon Detection," fig. 1, pp. 896, Copyright 2007, IEEE.]

12.4. Geiger mode operation

12.4.1. Basic theory.

When biased above the breakdown voltage, the avalanche photodiodes are capable of detecting single photons. This operation mode is called Geiger mode for analogy with the x-ray detection and the APDs that show this capability are called single-photon avalanche diodes (SPADs). This performance has positively contributed to the progress of numerous fields: basic quantum mechanics, cryptography, astronomy, single molecule detection, luminescence microscopy, fluorescent decays and luminescence in physics, chemistry, biology, and material science, diode laser characterization, optical fiber testing in communications and in sensor applications, laser ranging in space applications and in telemetry, and photon correlation techniques in laser velocimetry and dynamic light scattering [Cova et al. 1996]. Thus, integrated photon counting modules are today commercially available for applications such as time-correlated spectroscopy, quantum key distribution, astronomical observations, laser detection and ranging (LADAR) or bio-agent detection.

In order to reach those capabilities, the APD must be connected to a quenching circuit, which should be able to attenuate the avalanche multiplication and subsequent current increase after the photon arrival. The external circuitry could actively or passively respond to the avalanche. The simplest passive quenching circuit is a resistor connected in series with the APD (see Fig. 12.10). For this basic configuration, the detection process and avalanche quenching takes place as follows:

1. The photon is absorbed in the active volume of the avalanche photodiode while the device is negatively biased over the breakdown voltage.
2. Through consecutive multiplication events, the initial charge is amplified raising the output external current up to the milliamp range.
3. The voltage drop in the external load resistor (R_L) builds up causing the bias voltage on the APD to reduce below the breakdown voltage. The current delivered by the APD decreases consequently.
4. In a few nanoseconds, the flowing current becomes negligible, and the device is again biased over the breakdown voltage waiting for the arrival of a new photon.

The result of this sequence of events is a current pulse that can be observed by using an oscilloscope. SPAD timing jitter or timing resolution is defined as the width of the statistical distribution of the delay between the true arrival time of the photon at the sensor and the measured time marked by the output pulse current leading edge. This parameter improves as the overbias increases.

A discriminator circuit is needed in order to detect the pulse. The discriminator voltage will determine which pulses result in a count and which ones are neglected. Hence, it is value must be carefully adjusted in order to maximize the signal-to-noise ratio, i.e. the number of detected pulses over the number of undesirable counts. These undesirable counts are the result of the spurious emission of trapped charges that trigger the avalanche in absence of photons. They are called dark counts and must be minimized through material growth, device fabrication and optimum selection of operation parameters.

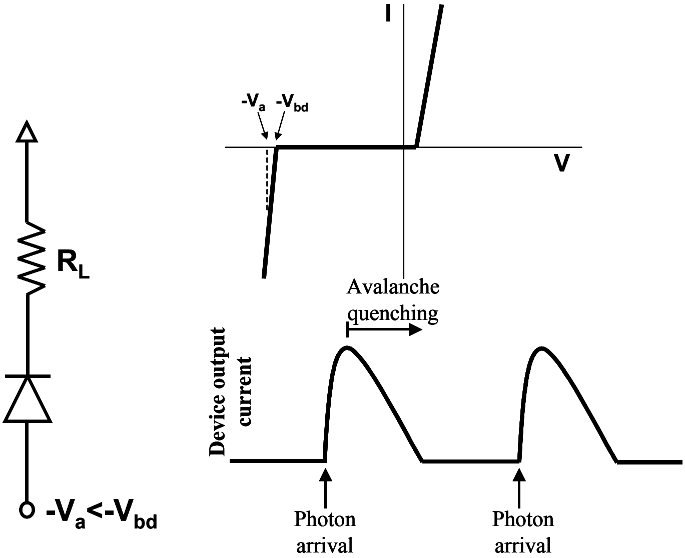


Fig. 12.10. Basic detection circuit for single-photon detection using APDs in Geiger mode.

The difference between the applied bias and the breakdown voltage is called overbias or excess bias. In practical terms, the probability of avalanche (PA), i.e. the probability that a carrier or electron-hole pair initiate a current pulse, will strongly depend on this parameter. Moreover, PA will depend on the position where the pair is generated, or the carrier injected, and on the ionization coefficients. In the most general case in which the electric field is not uniform throughout the depletion region, the ionization coefficients will be dependent on the position. Following McIntyre's arguments, the probability of avalanche when a hole is injected in the depletion region at $x = 0$ ($PA_h(0)$) can be calculated using

Eq. (12.14) $$PA_h(0) = 1 - \exp\left(-\int_0^W \beta(x) PA_p(x)\,dx\right)$$

where $PA_p(x)$ is the probability of avalanche for an electron-hole pair generated at x expressed as

Eq. (12.15) $$PA_p(x) = \frac{PA_h(0)f(x)}{PA_h(0)f(x) + 1 - PA_h(0)}$$

and where

Eq. (12.16) $$f(x) = \exp\left(\int_0^x (\alpha(x) - \beta(x))dx\right)$$

On the other hand, the probability of avalanche ($PA_e(W)$) for injection of electrons at the other side of the depletion region ($x = W$) can be determined from $PA_h(0)$ using

Eq. (12.17) $$\frac{PA_e(W)}{1 - PA_e(W)} = \frac{PA_h(0)}{1 - PA_h(0)} f(W)$$

These expressions are useful to predict the probability of avalanche (PA) in real designs in order to optimize the dark count probability and the single-photon detection efficiency. Fig. 12.11 shows an example of the use of these equations to calculate PA for hole injection at $x = 0$ in an InP-based multiplication region.

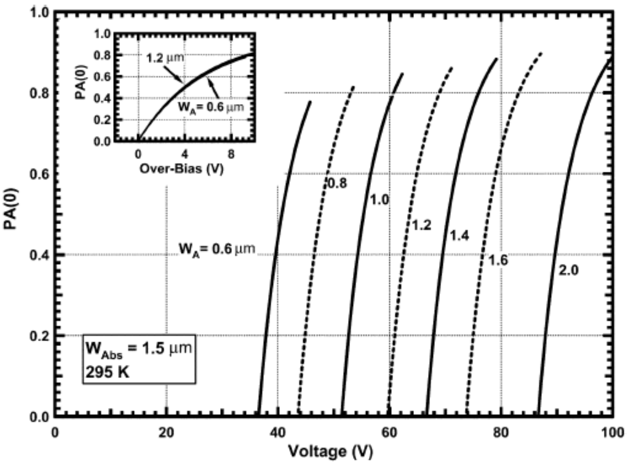


Fig. 12.11. Probability of avalanche for injection of holes at x = 0 calculated for InGaAsP/InP avalanche photodiodes with different multiplication region thicknesses (W_A). [Reproduced with permission from IEEE Journal of Quantum Electronics Vol. 42, No. 8, J.P. Donnelly, E.K. Duerr, K.A. McIntosh, E.A. Dauler, D.C. Oakley, S.H. Groves, C.J. Vineis, L.J. Mahoney, K.M. Molvar, P.I. Hopman, K.E. Jensen, G.M. Smith, S. Verghese, and D.C. Shaver, "Design Considerations for 1.06 μm InGaAsP–InP Geiger-Mode Avalanche Photodiodes," fig. 3, pp. 800, Copyright 2006, IEEE.]

Contrarily to the linear-mode operation of avalanche photodiodes, the impact of the excess noise factor on the Geiger-mode performance is less since the electrical pulses are converted into digital signals of the same amplitude after the discriminator circuit. This property makes possible to reach the so-called zero-read noise performance in absence of dark counts.

The multiplication gain loses part of its meaning in this operation mode because the photocurrent or device response is modulated by the response of the external circuit. An "effective gain" (M_G) can be calculated from the peak current of the pulse delivered by the APD (I_{peak}) as

$$\text{Eq. (12.18)} \quad M_G = \frac{I_{peak}}{I_{0,ph}} = \frac{I_{peak}}{P_{in}} \frac{hc}{q\lambda\eta}$$

but its value is not an intrinsic parameter of the device as in linear-mode APDs since it is dependent on the external circuit characteristics.

12.4.2. Passive avalanche quenching

The SPAD are passively quenched when the avalanche current quenches itself through a load resistor (R_L), as in the example above. This load resistor has typical values between 50 kΩ and 500 kΩ. The resistor connected to ground (R_s) provides matched termination for a 50 Ω coaxial cable. It is common to connect the output of this simple circuit to a comparator that produces a standard logic signal for pulse counting and timing.

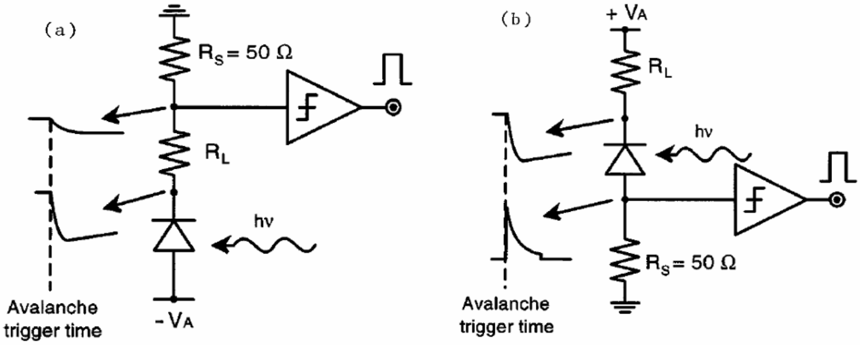


Fig. 12.12. Basic configurations of passive quenching circuit: (a) voltage-mode output and (b) current-mode output.[Adapted with permission from Applied Optics Vol. 35, No. 12, S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," fig. 4, pg.1960, Copyright 1996, Optical Society of America (OSA).]

This passive quenching circuit has two possible configurations: voltage-mode output Fig. 12.12a) and current-mode output (Fig. 12.12b). The former provides longer pulses, which might be convenient to visualize them in the oscilloscope but might hinder high-speed detection. On the contrary, the configuration with current-mode output allows high detection rates.

12.4.3. Active avalanche quenching

An alternative to passive quenching is active quenching. This can be implemented by forcing the avalanche quenching using an external circuit that suddenly reduces the APD bias when a current pulse is detected. It is noticeable that the amplitude of the quenching pulse must be equal or higher than the overbias. The external circuit can be a pulse-booster circuit or a combination of electronic switches and external DC power supplies.

The main purpose of this mode is to avoid slow recovery from avalanche pulses in the passive mode, which is common in certain materials. However, the design of passive quenching circuits involves more components and is more delicate than the passive quenching. Circuit oscillations and spurious retriggering of the external circuit must be avoided. In the example illustrated in Fig. 12.13, the network in the dotted box compensates the current pulses injected by the quenching pulse through the SPAD capacitance, thus avoiding circuit oscillation. The capacitor value must be chosen to match the detector capacitance.

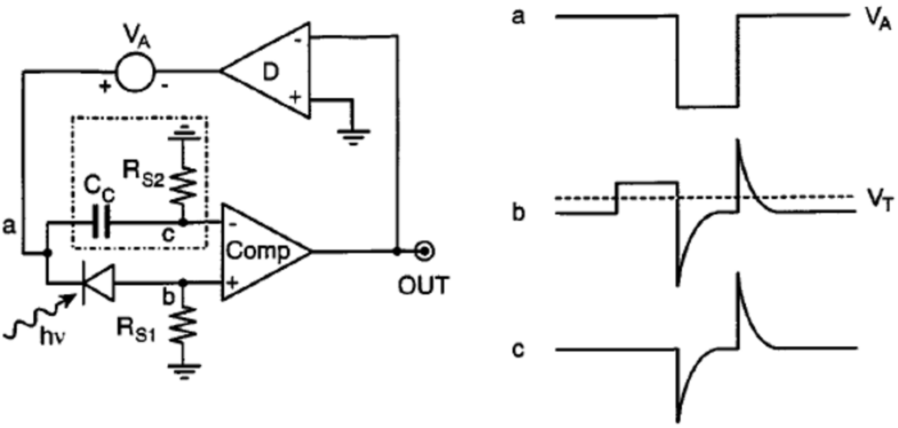


Fig. 12.13. Simplified diagram of active quenching electronics. The voltage waveforms depicted correspond to the circuit nodes marked with the same letter. [Adapted with permission from Applied Optics Vol. 35, No. 12, S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," fig. 10, pg.1965, Copyright 1996, Optical Society of America (OSA).]

12.4.4. Gated detection

The use of gated operation is recommended when the application requires detecting photons only during a short interval or when the effect of trapping should be minimized. The gated detection in SPADs is implemented by biasing the device at a DC voltage (V_{DC}) slightly lower than the breakdown voltage, and applying a pulsed signal that repeatedly drives the device above the breakdown voltage only during the pulse width. The device is only capable to detect photons during that short time, being insensitive to the photon arrival during the rest of the period.

Applications like time-correlated fluorescence or laser ranging require gated operation to avoid the impact of a strong excitation pulse that precedes the photon emission to be detected. Furthermore, the gate can be scanned along a period to plot the photon emission probability as a function of time, as sketched in Fig. 12.14. This technique called Time Correlated Single-Photon Counting is very useful in determining emission lifetimes. Quantum cryptography can also use gated detection systems since the transmission frequency between sender and receiver can be set beforehand.

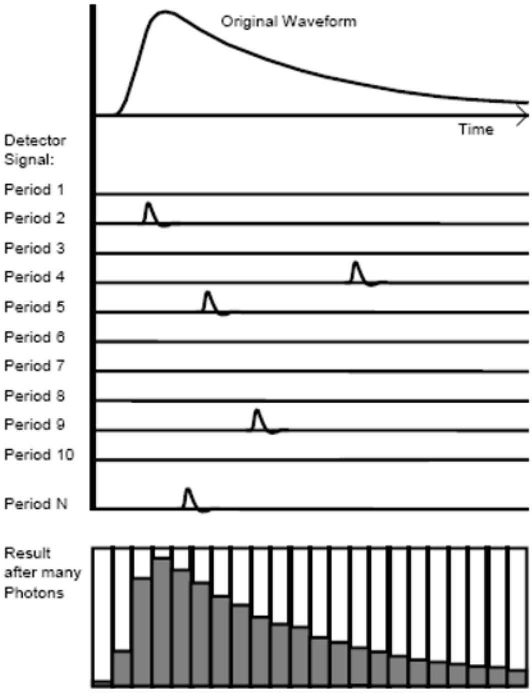

Original Waveform
Time
Detector Signal:
Period 1
Period 2
Period 3
Period 4
Period 5
Period 6
Period 7
Period 8
Period 9
Period 10
Period N
Result after many Photons

Fig. 12.14. Time correlated single-photon counting measurement principle.[Reproduced with permission from Boston Electronics Corporation Notes, "What is Time Correlated Single-Photon Counting," fig. 1, pg. 3, Copyright 2008, Boston Electronics Corporation]

On the other hand, this operation mode is effective in avoiding that the emission of charge from traps during the off-intervals triggers the avalanche in absence of photons and raises the dark count rate. For that aim, it is needed that the time that the gate remains off is longer than the trap release time. Otherwise, the released charge could trigger an avalanche pulse during the next detection gate. This phenomenon is called after-pulsing and will be discussed in more detail in the next section.

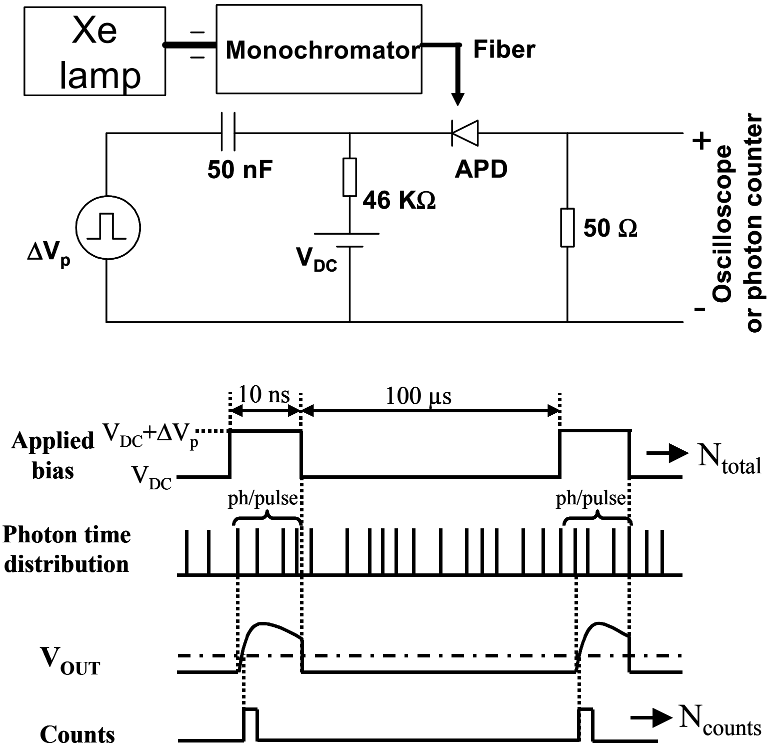

Xe lamp
Monochromator
Fiber
50 nF
APD
46 KΩ
50 Ω
ΔV_p
V_DC
+
-
Oscilloscope or photon counter
10 ns
100 µs
Applied bias
V_DC+ΔV_p
V_DC
ph/pulse
ph/pulse
N_total
Photon time distribution
V_OUT
Counts
N_counts

Fig. 12.15. Top: Gated quenching circuit. Bottom: Schematic of the photon detection process in gated systems.

Fig. 12.15 shows a typical circuit for the operation in gated mode. The pulse needed to bias the device over the breakdown voltage is usually provided by a pulse generator. Either a bias-T or a capacitor and a resistor are required to connect the DC power supply and the pulse generator to the APD input. The light source is attenuated to attain a photon flux lower than about one photon per pulse. The signal is measured as the voltage drop through a 50 Ω load resistor. Unless the pulse width is too large, additional quenching circuit is not required since the end of the pulse drastically reduces the diode voltage under the breakdown voltage attenuating the avalanche current.

The output of the circuit is connected to an oscilloscope or a photon counter. Every time the voltage across the 50 Ω resistor overcomes the threshold voltage, the system count up one. At the end of the measurement time, the system would have registered N_{counts}, a number necessarily lower than the total number of bias pulses N_{total} generated to bias the APD during that time. The pulse detection efficiency (PDE) is defined as

$$\text{Eq. (12.19)} \quad PDE\ (\%) = 100 \times \frac{N_{counts}}{N_{total}}$$

turns to be the well-known single-photon detection efficiency (*SPDE*) when the photon flux is made equal to one photon per pulse. This parameter is one of the figures-of-merit of the SPADs and enables comparison between them. As illustrated in Fig. 12.16, at high photon fluxes, the PDE is 100% and approaches the SPDE value as we reduce the number of photons per pulse.

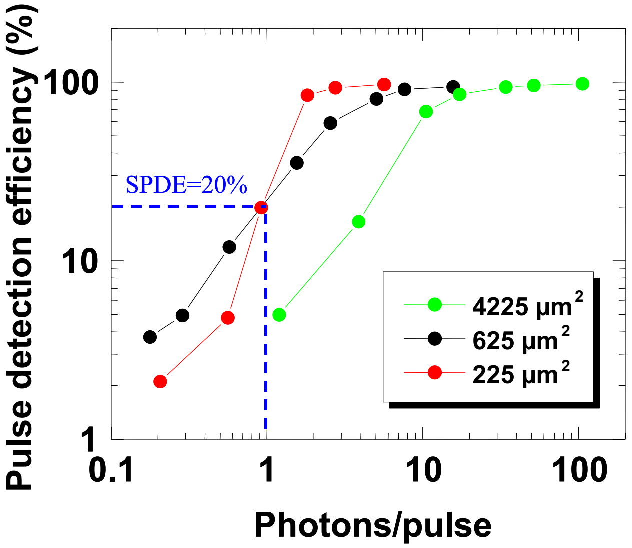


Fig. 12.16. Pulse detection efficiency as a function of the photon flux for various size GaN APDs. [Reproduced with permission from Applied Physics Letters Vol. 91, J.L. Pau, R. McClintock, K. Minder, C. Bayram, P. Kung, M. Razeghi, E. Muñoz, and D. Silversmith, "Geiger-mode operation of back-illuminated GaN avalanche photodiodes," fig. 3, pg. 041104-2, Copyright 2007, American Institute of Physics (AIP)]

Unfortunately, when the measurements are performed in the dark, the probability to register pulses and consequently counts is not negligible. As we discussed before, the origin of these pulses is the charge release from deep levels or traps in the semiconductor. These charges can trigger the avalanche in absence of photons when the bias pulse arrives at the SPAD. Therefore, during a measurement time in darkness, the system will register a certain number of counts N_{counts}^{dark} out of the total number of pulses N_{total}. The dark count probability is defined as the ratio between both magnitudes ($N_{counts}^{dark} / N_{total}$). This is another important figure-of-merit to minimize in the device optimization. For the sake of comparison between different measurement conditions, the dark count probability is often normalized by the pulse width. The resultant parameter is called dark count rate (*DCR*) and is therefore expressed as

$$\text{Eq. (12.20)} \quad DCR = \frac{N_{counts}^{dark}}{\Delta t_p N_{total}} \quad [\boldsymbol{Hz}]$$

In general, the overbias increase results in a *SPDE* enhancement but also in a higher *DCR*. For comparison purposes, it is sometimes useful to unify both parameters to optimize device performance. As in standard detectors, SPDE and DCR can be combined in a single figure-of-merit: the noise equivalent power (NEP). In SPADs, this is calculated as

$$\text{Eq. (12.21)} \quad NEP = \frac{h\nu}{SPDE}\sqrt{2\,DCR} \quad [\boldsymbol{W \cdot Hz^{-1/2}}]$$

Alternatively, one can use the detectivity corrected by the detector area (*D**) to compare devices with different active areas. This can be obtained from the NEP using

$$\text{Eq. (12.22)} \quad D^* = \frac{\sqrt{A}}{NEP} \quad [\boldsymbol{W^{-1} \cdot cm \cdot Hz^{1/2}}]$$

12.4.5. Device limitations

One of the major limitations of SPADs is their non-capability to resolve the number of photons that arrive in one single pulse. The reason is that while the avalanche is active the detector is disabled to detect new photons. The amplitude of the current pulse delivered by the APD is quite independent on the number of photons, which impedes to detect bunches of two or three photons. As the photon emission of lasers and LEDs follows a Poissonian distribution, it is not unlikely to find several photons impinging the diode within the same pulse. This is a problem in quantum cryptography, where photon bunching could spoil some of the advantages of the quantum communication. One of the major scientific challenges to overcome this drawback is to develop truly single-photon emitters that can supply single photons on-demand without bunching.

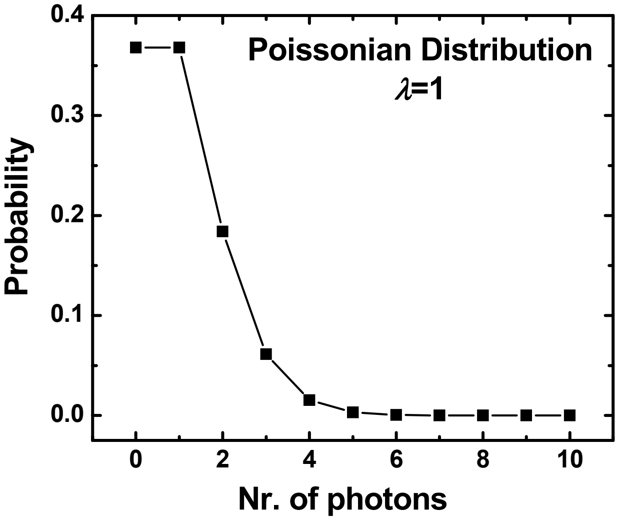


Fig. 12.17. Poissonian distribution followed by the photon emission statistics in an LED or laser diode for an average photon flux of one ($\lambda = 1$). It is noticeable that the probability to get none, two, three or four photons per pulse is not negligible.

12.4.6. After-pulsing

Impurities, dislocations, stacking faults, interstitial atoms, vacancies, substitutional atoms all are crystal defects that can give rise to deep levels in the semiconductor bandgap. The presence of those levels in the multiplication region of the avalanche photodiodes can have dramatic consequences. The main problem is that they can trap charge during an avalanche event. The charge remains in the traps a characteristic time known as *detrapping time*. When that charge is released, there is a high probability to trigger a new avalanche, which results in a pulse that might be detected by the discriminator circuit but unrelated with the photon arrival. This situation makes necessary to disable either the APD or the detection of the discriminator circuit for a specific time called *dead time* to avoid registering the event. Its value must be larger than the *detrapping time* in order to ensure that all the traps have been emptied before enable the system back. In gated detection, one way to fulfill this requirement is to make the inverse of the pulse rate larger than the *detrapping time*.

As the charge emission from the traps is a thermally activated process, the *detrapping time* would be larger as temperature decreases (see Fig. 12.18). This problem seriously limits the maximum operation frequency of these devices and, consequently, the transmission rate in quantum communication systems. In InGaAs/InP photodiodes for telecom wavelengths, it limits the usable frequency to a few megahertz.

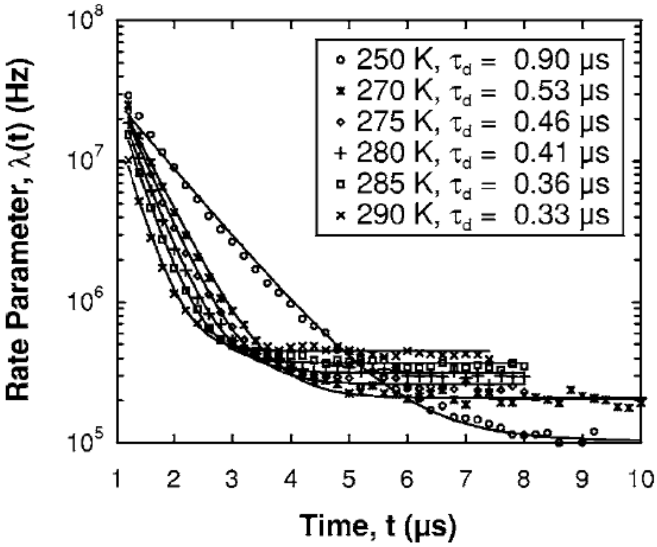


Fig. 12.18. Rate (λ) of subsequent avalanche events measured at different delay times (t) after the first avalanche. The calculated detrapping times (τ_d) at different temperatures are shown in the inset. [Reproduced with permission from Applied Physics Letters Vol. 88, K.E. Jensen, P.I. Hopman, E.K. Duerr, E.A. Dauler, J.P. Donnelly, S.H. Groves, L.J. Mahoney, K.A. McIntosh, K.M. Molvar, A. Napoleone, D.C. Oakley, S. Verghese, C.J. Vineis, and R.D. Younger, "Afterpulsing in Geiger-mode avalanche photodiodes for 1.06 µm wavelength," fig.2, pg. 133503-2, Copyright 2006, American Institute of Physics (AIP.]

12.5. Examples of single-photon avalanche photodiodes

12.5.1. Silicon single-photon avalanche diodes

Commercial single-photon counting modules based on Si APDs are available since some years ago. In these modules, the detectors are usually mounted on thermoelectric coolers to keep the temperature a few tens of degrees below 0 °C to reduce the dark count rate. Detector diameters of about 200 µm and dark count rates lower than 100 counts/sec are typical in high-performance modules. Due to the high crystal quality and well known processing techniques in silicon, these modules present dead times of about 50 ns with an after-pulsing probability lower than 0.5%. Furthermore, gating is not required, which simplifies its use and makes them available for a wide range of applications. Their single-photon detection efficiency onsets at 400 nm (5%), peaks at 650 nm (65%) and extends to the near-IR, but rarely presents operational values at wavelengths higher than 1060 nm (2%). Thus, although Si single photon avalanche diodes (SPADs) outperform photocathodes for certain applications in the 400–1060 nm range, their low efficiency in the ultraviolet (UV) and at telecom wavelengths makes necessary to seek other technologies for those ranges.

A technique called *up-conversion* is used to expand the detection range of the Si detectors beyond 1 µm. The basic idea is to use a periodically-poled lithium niobate (PPLN) waveguide pumped by an intense escort laser pulse (ω_e) to convert input IR photons (ω_i) to visible photons (ω_o), for which Si SPADs present their highest efficiency. Lithium niobate is a non-linear ferroelectric material whose domains point alternatively to opposite directions in the waveguide thanks to the use of periodically structured electrodes. By properly selecting the poling period of the crystal, we can create a quasi-phase matching situation, which allows us to take advantage of the non-linear properties to achieve conversion. In that case, energy conservation requires that

Eq. (12.23) $\omega_o = \omega_i + \omega_e$.

The experimental set-up is shown in Fig. 12.19. An attenuated IR laser at 1.55 µm is combined with a high-power pump laser in a wavelength division multiplexer (WDM) before injection into the PPLN. At its output, the separation of the up-converted signal, the pump signal and the spurious light is carried out using filters, a prism, and a spatial filter (SM). Finally, the up-converted signal is focused onto the single-photon counting module (SPCM).

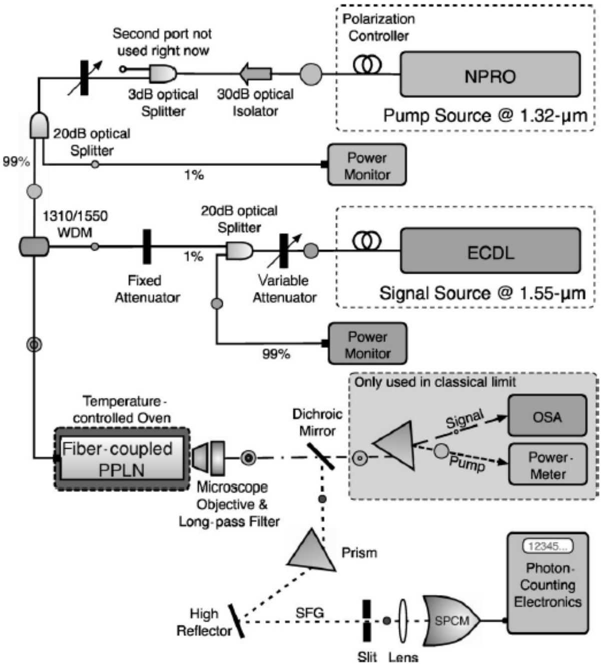

Polarization Controller
Second port not used right now
NPRO
3dB optical Splitter
30dB optical Isolator
Pump Source @ 1.32-µm
20dB optical Splitter
99%
Power Monitor
1%
1310/1550 WDM
20dB optical Splitter
ECDL
1%
Fixed Attenuator
Variable Attenuator
Signal Source @ 1.55-µm
Power Monitor
99%
Only used in classical limit
Temperature-controlled Oven
Dichroic Mirror
Signal
OSA
Fiber-coupled PPLN
Pump
Power-Meter
Microscope Objective & Long-pass Filter
Prism
12345...
Photon-Counting Electronics
High Reflector
SFG
SPCM
Slit
Lens

Fig. 12.19. Experimental setup for single-photon detection at 1.55 µm using up-conversion. OSA, optical spectrum analyzer. [Reprinted with permission from Optics Letters Vol. 30, C. Langrock, E. Diamanti, R.V. Roussev, Y. Yamamoto, M.M. Fejer, and H. Takesue, "Highly efficient single-photon detection at communication wavelengths by use of upconversion in reverse-proton-exchanged periodically poled LiNbO$_3$ waveguides," fig. 1, pp. 1726, Copyright 2005,Optical Society of America (OSA).]

12.5.2. InGaAs/InP single-photon avalanche diodes

Due to the limiting properties of Si to detect telecom wavelengths, many researchers focused their attention on the Geiger mode operation of InGaAs/InP APDs to enable quantum cryptography, or better called, quantum key distribution in communication systems. This technique allows two parties to share a random bit string known only to them, which can be used as a key to encrypt and decrypt messages. A unique property of these systems is the ability of the two communicating users to detect the presence of any third party trying to gain knowledge of the key. In the last years, the progress in developing InGaAs/InP SPADs has made possible the fabrication of the first quantum cryptography prototypes operating over distances extending to several dozens of kilometers.

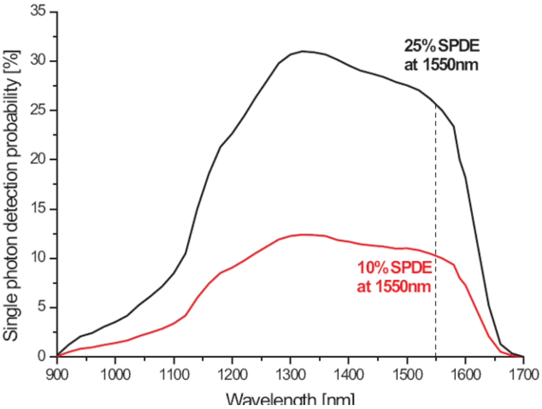


Fig. 12.20. Single-photon detection probability versus wavelength in commercial InGaAs/InP SPADs for two different overbias values [Reprinted with permission from www.idquantique.com, "Single-Photon Detector Module Id201 Datasheet," fig. 1,pg.3, Copyright 2008, IdQuantique, Switzerland].

Commercial InGaAs/InP SPADs present single-photon detection efficiencies between 10 and 25% at telecom wavelengths, outperforming Si counterparts and photocathodes at those wavelengths. Below 1 µm, their efficiency drops drastically, preventing their use in the visible range. They operate in gated mode and are thermoelectrically cooled at temperatures around −50 °C to reduce the dark count probability to $1–4 \times 10^{-4}$ per 1 ns gate (DCR = 100–400 kHz). In contrast to Si SPADs, dead times in the µs-range are used in these devices with afterpulsing probabilities of $3–6 \times 10^{-3}$. These characteristics hinder their operation at frequencies higher than 1 MHz, despite timing jitters of only a few cents of ps are obtained.

High performance devices have been recently reported with single-photon detection efficiencies over 30% and dark count rates close to 10 kHz at around 200 K. They usually rely on separate-absorption-charge-multiplication APDs with planar structure as the one shown in Section 12.3.2.

12.5.3. GaN single-photon avalanche diodes

Due to their narrow bandgap, silicon SPADs show poor single-photon detection efficiencies in the UV. In particular, surface dead layers and phonon emission during carrier relaxation significantly reduce the probability of detecting single photons. Therefore, wide bandgap materials present intrinsic advantages such as higher efficiency and lower noise to detect UV photons.

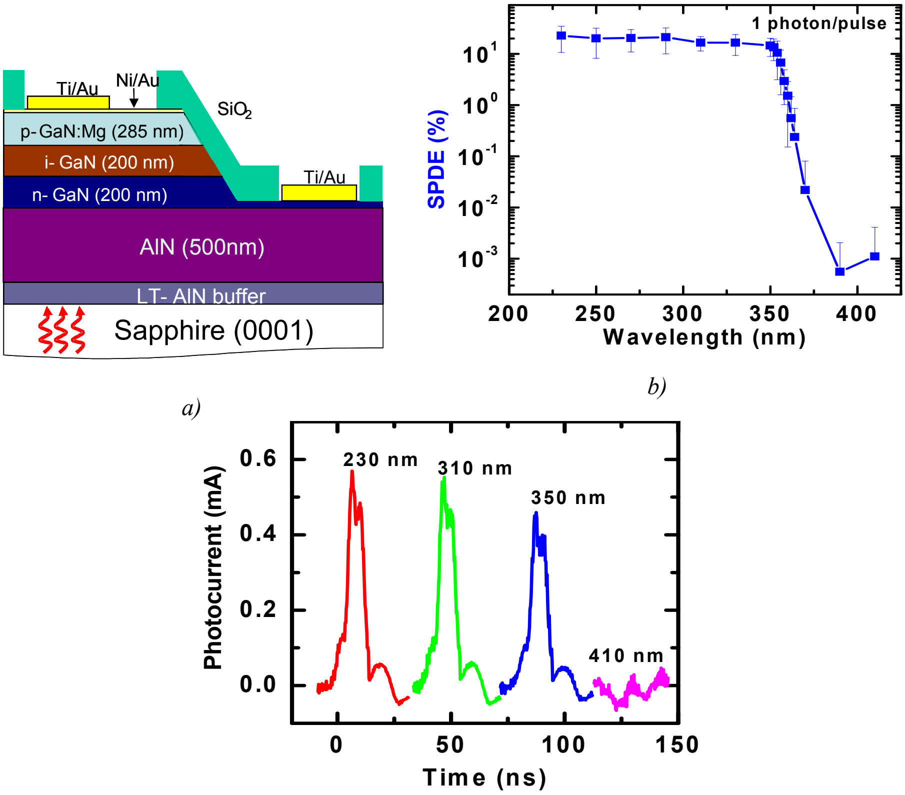
GaN APDs present excellent properties for UV sensing. In combination with AlGaN ternary alloy, the cut-off wavelength can be tailored in a wide spectral range from 363 nm to 200 nm. This capability, together with other properties such as high radiation hardness, good chemical stability, high *n*- and *p*-type doping levels, and the possibility to form heterostructures, makes these devices good candidates to satisfy the needs of many civilian and military UV applications.

One example of these applications is the fast and reliable detection of threat agents such as anthrax, smallpox or plague bio-molecules. The idea is to combine UV and visible SPADs to build bio-fluorescence sensors capable to continuously monitor the air quality. Benefiting from the distinguishable signatures of the threat agents on the ratio between the visible and UV signals, their presence can be detected.

The ability of the ozone layer to filter out the solar radiation in the UV-C band (200–280 nm) makes the photodetectors with cut-off wavelengths below 280 nm intrinsically insensitive to the sunlight. For that reason, these detectors receive the adjective of "solar-blind" and open possibilities for new applications. Thanks to the fact that single photon detection prevents eavesdropping through quantum encryption, ultra-secure communication systems based on solar-blind SPADs have been envisaged for non-line-of-sight and inter-satellite communications.

High-performance GaN APDs consist of *p-i-n* homojunctions grown on top of UV-transparent AlN templates (Fig. 12.21a). As holes have demonstrated to have higher ionization coefficient than electrons in this material, these structures are specifically designed for back-illumination to promote the injection of holes into the multiplication region, to minimize noise and to maximize multiplication gain.

Although still in an early stage of development, single-photon detection capabilities have been demonstrated in gated mode between 360 and 200 nm (Fig. 12.21b). SPDEs between 20 and 30% have been achieved in this range with dark count rates lower than 10 kHz in small area devices (<625 μm^2). Experiments show very low signals above the cut-off wavelength ($\lambda = 363$ nm) with no counts at wavelengths higher than 410 nm (Fig. 12.21c). Compared to narrow-bandgap technologies, this is an additional advantage for many UV applications since the need of external filtering to block the visible and IR radiation is considerably alleviated.



a)

b)

c)

Fig. 12.21. a) Back-illuminated GaN APD structure; b) single-photon detection efficiency as a function of wavelength; c) current pulses delivered under illumination with UV light (230, 310, and 350 nm) and visible light (410 nm). [Adapted with permission from Applied Physics Letters Vol. 91, J.L. Pau, R. McClintock, K. Minder, C. Bayram, P. Kung, M. Razeghi, E. Muñoz, and D. Silversmith, "Geiger-mode operation of back-illuminated GaN avalanche photodiodes," fig. 1,2 & 3, pg. 041104-1 & 041104-2, Copyright 2007, American Institute of Physics (AIP).]

12.6. Summary

In this Chapter, we started by making an historical review of the development of avalanche photodetectors for single-photon detection. A brief introduction to the theory of avalanche photodetectors and fabrication techniques was provided and typical strategies to improve performance such as guard rings, beveled mesa walls or recessed windows are presented. Useful expressions for the calculation of multiplication gain and noise characteristics as a function of the ionization coefficients are given.

In Section 12.4, the principles of operation of SPADs in Geiger mode are described along with the types of quenching circuitry. The probability of detecting a photon is ruled by the probability of avalanche whose formulation is revisited for device modeling purposes. The concepts of gating and after-pulsing in this mode are presented. Finally, the state-of-the-art of SPADs in Si, InGaAs/InP and GaN is reviewed. These devices are spectrally complementary and satisfy most of the needs for single-photon detection from the UV to the near-IR, including telecom wavelengths.

References

Bai, X., McIntosh, D., Liu, H. and Campbell, J.C., “Ultraviolet single photon detection with Geiger-mode 4H-SiC avalanche photodiodes,” *IEEE Photonic Technology Letters* 19, pp. 1822-1824, 2007.

Baraff, G.A., “Distribution functions and ionization rates for hot electrons in semiconductors,” *Physical Review* 128, pp. 2507-2517, 1962.

Bennett, C.H., Bessett, F., Brassard, G., Salvail, L. and Smolin, J., “Experimental Quantum Cryptography”, *Journal of Cryptography* 5, pp. 3-28, 1992.

Chynoweth, A.G. and McKay, K.G., “Photon emission from avalanche breakdown in silicon,” *Physical Review* 102, pp. 369-376, 1956.

Cova, S., Ghioni, M., Lacaita, A., Samori, C. and Zappa, F., “Avalanche photodiodes and quenching circuits for single-photon detection,” *Applied Optics* 35, pp. 1956-1976, 1996.

Daudet H., Deschamps P., Dion B., MacGregor A.D., MacSween D., McIntyre R.J., Trottier C. and Webb P.P., “Photon counting techniques with silicon avalanche photodiodes,” *Applied Optics* 32, pp. 3894-3900, 1993.

Gullikson E.M., Gramsch E. and Szawlowski M., “Large-area avalanche photodiodes for the detection of soft x-rays,” *Applied Optics* 34, pp. 4662-4668, 1995.

Hiskett, P.A., Smith, J.M., Buller, G.S. and Townsend, P.D., “Low-noise single-photon detection at wavelength 1.55 µm,” *Electronics Letters* 37, pp. 1081 (2001).

Liu, Y., Forrest, S.R., Hladky, J., Lange, M.J., Olsen, G.H. and Ackley, D.E., “A Planar InP/InGaAs Avalanche Photodiode with Floating Guard Ring and Double Diffused Junction,” *Journal of Lightwave Technology* 10, pp. 182-193, 1992.

McClintock, R., Pau, J.L., Minder, K., Bayram, C., Kung, P. and Razeghi, M., “Hole-initiated multiplication in back-illuminated GaN avalanche photodiodes,” *Applied Physics Letters* 90, p. 141112, 2007.

McIntyre, R.J., “Multiplication noise in uniform avalanche diodes,” *IEEE Transactions on Electron Devices* ED-13, pp. 164-168, 1966.

Minder, K, Pau, J.L., McClintock, R., Kung, P., Bayram, C. and Razeghi, M., “Scaling in back-illuminated GaN avalanche photodiodes,” *Applied Physics Letters* 91, p. 073513, 2007.

Nightingale, N.S., “A new silicon avalanche photodiode photon counting detector for astronomy,” *Exploratory Astronomy* 1, pp. 407-422, 1991.

Pau, J.L., McClintock, R., Minder, K., Bayram, C., Kung, P., Razeghi, M., Muñoz, E. and Silversmith, D., "Geiger-mode operation of back-illuminated GaN avalanche photodiodes," *Applied Physics Letters* 91, p. 041104, 2007

Shockley, W., "Problems related top-n junctions in silicon," *Czechosolvak Journal of Physics* B11, pp. 81-121, 1961.

Webb, P.P. and Mclntyre, R.J., "Single Photon Detection with Avalanche Photodiodes," *Bulletin of the American Physical Society* 15, p. 813, 1970.

Wolff, P.A., "Theory of Electron Multiplication in Silicon and Germanium," *Physical Review* 95, pp. 1415-1420, 1954.

Further reading

Cova, S., Ghioni, M., Lacaita, A., Samori, C. and Zappa, F., "Avalanche photodiodes and quenching circuits for single-photon detection," Applied Optics 35, pp. 1956-1976, 1996.

Donnelly, J.P., Duerr, E.K., McIntosh, K.A., Dauler, E.A., Oakley, D.C., Groves, S.H., Vineis, C.J., Mahoney, L.J., Molvar, K.M., Hopman, P.I., Jensen, K.E., Smith, G.M., Verghese, S. and Shaver, D.C., "Design Considerations for 1.06-µm InGaAsP-InP Geiger-Mode Avalanche Photodiodes," IEEE Journal of Quantum Electronics 42, pp. 797-809, 2006.

Stillman, G.E. and Wolfe, C.M., "Avalanche photodiodes", *Semiconductors and Semimetals, vol. 12: Infrared Detectors II*, R.K. Willardson and A.C. Beer, eds., Academic, New York, 1977.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

Problems

1. Knowing that the inverse of the multiplication coefficients can be generally formulated as:

$$M_n^{-1} = 1 - \int_0^W \alpha \exp\left[-\int_0^x (\alpha - \beta)\, dx'\right] dx$$

$$M_p^{-1} = 1 - \int_0^W \beta \exp\left[-\int_x^W (\alpha - \beta)\, dx'\right] dx\,,$$

demonstrate that both expressions can be rewritten as:

$$M_n = \frac{[1-(\beta/\alpha)]\exp\{\alpha W[1-(\beta/\alpha)]\}}{1-(\beta/\alpha)\exp\{\alpha W[1-(\beta/\alpha)]\}}$$

$$M_p = \frac{[1-(\beta/\alpha)]}{1-(\beta/\alpha)\exp\{\alpha W[1-(\beta/\alpha)]\}}$$

when the electric field is uniform, i.e. ionization coefficients are constant, across the multiplication region.

2. Calculate the maximum photocurrent expected from the illumination of an APD with an optical power of 1 nW at λ = 1 μm, knowing that the quantum efficiency is 60% and the multiplication gain is equal 100. Compare that value with the photocurrent obtained under the same circumstances at λ = 400 nm.

3. What is the spectral noise density per nA in an APD at a 100 gain when $k = 0.04$ under electron injection? And under hole injection?

4. Calculate the probability of avalanche for an electron injected at x = 0, considering that the ionization coefficients α, β are equal to 3×10^4 and $10^4\,\mathrm{cm}^{-1}$, respectively, and the electric field is constant across the multiplication region thickness ($W = 500$ nm).

5. A single-photon detection experiment is carried out in gated mode at a pulse repetition rate of 10 kHz and a pulse width of 10 ns. After illumination with one photon (λ = 700 nm) per pulse, 2,153 counts are registered in the discriminator circuit. Taking into account that the same measurement performed in the dark delivered 10 counts, determine the detectivity of the SPAD under those conditions.

6. What is the wavelength needed to pump a PPLN crystal in an up-conversion experiment if we aim to convert 1.5 μm photons into 650 nm photons.

13. Terahertz Device Technology

13.1. Introduction
13.2. Applications
  13.2.1. THz spectroscopy
  13.2.2. T-ray imaging
  13.2.3. THz research tool
13.3. Broadband terahertz sources
13.4. Narrow band terahertz sources
  13.4.1. Optical converter
  13.4.2. Optically pumped gas lasers
  13.4.3. Semiconductor sources based on Si and Ge
13.5. Quantum cascade terahertz sources
  13.5.1. GaAs based terahertz QCLs
  13.5.2. InP based terahertz QCLs
13.6. Magnetic field effects
13.7. Difference frequency generation
13.8. GaN QCLs for high temperature operation
13.9. Summary

13.1. Introduction

The infrared region is a range of the electromagnetic spectrum split into three parts, the near-infrared (NIR), the mid-infrared (MIR) and the far-infrared (FIR). Previously, it was usual to refer to the terahertz (THz) range as the range of frequencies below FIR. Nowadays, the THz range is attributed to the electromagnetic spectrum which spans the FIR to the millimeter wave. This THz range of the electromagnetic spectrum starts at 1000 µm and finishes at 50 µm (0.3 THz to 6 THz). So, these wavelengths are commonly named THz, far-infrared or submillimeter wave. THz waves are located between the optical region (infrared) and the microwaves. Only a few compact, easy to use, room temperature sources are available. The lack of detectors in this spectral range is also noteworthy. For example, the most

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1_13,
© Springer Science+Business Media, LLC 2010

sensitive FIR detector is the bolometer which needs to be cooled down to liquid Helium temperatures.

13.2. Applications

Far-infrared (FIR) range is an area of the electromagnetic spectra which has lots of applications but it suffers from the lack of simple working devices which can emit and detect THz radiation. In this section, we present a brief overview of the many conceivable applications in this range. The applications for the THz can be found in astronomy and space research, biology imaging, security, industrial inspection, etc. Most of them rely on THz spectroscopy, which is in frequency domain, and/or THz imaging, which is in space and time domain. In addition, THz technology also find application in scientific researches, such as studies of photonic crystal and quantum cascade lasers.

13.2.1. THz spectroscopy

THz spectroscopy deals with the frequency response in the THz range. For a few examples, we present applications in astronomy and biology.

The applications of THz range in astronomy and space studies are very important. Like all human beings, we would like to know the origin of our star: the Sun and as a consequence the Earth. The birth of galaxies and new stars is supposed to be due to a gas cooling process which allows interstellar dust clouds to collapse. Typical temperatures of this dense interstellar gas are in a range from about 10 K in the cooler regions to hundreds of Kelvin in the hotter and usually denser parts [Phillips et al. 1992]. The corresponding frequencies ($kT \sim h\nu$) range from about 200 GHz to 4 THz. As a consequence, 98% of post Big-Bang photons in a typical galaxy lie in the FIR spectral region [Siegel et al. 2002]. Fig. 13.1 presents a schematic of the gas spectrum of emission and is dominated by some submillimeter line emission due to the rotation and vibration spectrum of diatomic hybrid molecules. Light molecule spectroscopy and atomic fine-structure spectroscopy are the key experiments to study these clouds. For example, the strongest emission of heavy molecules is by far that of CO because of the abundance of this species. The carbon line at 1.9 THz is a single spectral line responsible for 0.25% of the energy radiated by a galaxy [Siegel et al. 2002]. The pieces of information on atomic and ionized carbon, as well as on molecules like H_2O and O_2, help to understand the birth dynamics of new and old galaxies, stars and planets (Fig. 13.1). As we just pointed out, the interstellar medium is rich in chemical details. The three most abundant elements observed are the hydrogen, the oxygen and the carbon. The sub millimeter spectrum of the Orion cloud contains 30 different identifiable molecules from CO to CH_3OCH_3 [Phillips et al. 1992]. The specifics of each molecule, for example the electric dipole moment of CO and the magnetic dipole of molecular oxygen, allow transitions from the rotational and spin levels, giving rise to different light transitions. Utilizing this emitted light from the vibration and rotation spectra of molecules, opens up a range of applications on Earth. For example, the study of smog levels, CFC-gas level and ozone layer can be a relevant way to monitor the global warming and to give reliable figures on gas concentrations.

Thanks to FIR spectra, researchers in atmospheric studies may develop a better understanding of atmospheric dynamics, and acquire data to build realistic models and provide real-time feedback on atmospheric conditions. We note that the FIR has particular applications in the space environment because of the absorption of THz emission by the atmosphere. Communications between inter and intra satellite can be achieved in this spectral range without any interferences and eaves droppings from the Earth within the huge bandwidth offered by the THz.

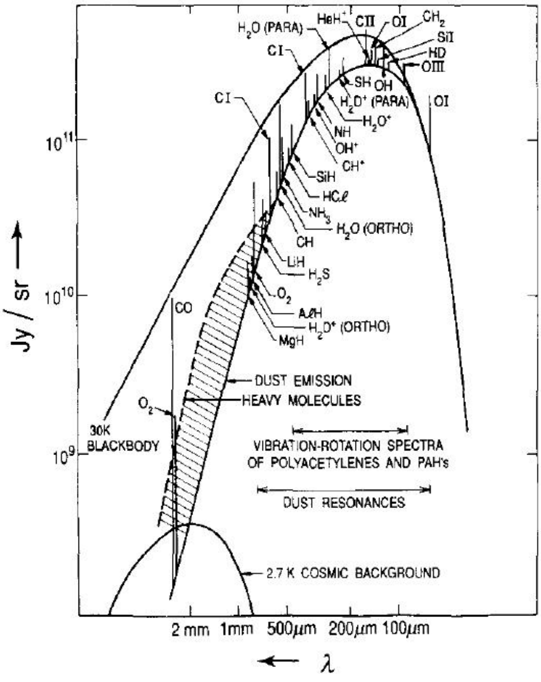


Fig. 13.1. A schematic presentation of some of the spectral content in the submillimeter band for an interstellar cloud. The spectrum includes dust continuum, molecular rotation line and atomic fine-structure line emissions. [Reprinted with permission from Proceedings of the IEEE Vol. 80, no. 11, Phillips, T.G. and Keene, J.," Submillimeter astronomy {heterodyne spectroscopy}," fig. 1, p. 1663. Copyright 1992, IEEE.]

Also on earth, applications to biology are very important and achievable in the submillimeter range. Using THz frequency spectroscopy, biological agents may be detected. Woolard et al. [2003] realized a differential absorption radar to detect the presence of *Bacillus Subtillus* spores. In this experiment, the level of biological hazard was monitored. As a consequence, this technique could be a powerful tool for evaluating bio-agent attacks. The THz spectroscopic tool has also been used to study some deoxyribonucleic acid (DNA) molecules. Detection of single base mutations of molecules has been demonstrated [Nagel et al. 2002]. As we know, the evolution of species, or some genetic diseases, come from genetic mutations. This field of research is relevant and needs powerful and easy to use THz sources and detectors. With the same experimental technique of THz transmission analysis, different absorption features from DNA samples have been explained by the presence of specific multiple dielectric resonances in the submillimeter range [Brucherseifer et al. 2000] [Woolard et al. 2002]. Along the way, intrinsic properties of the particular DNA sequence have been demonstrated. As a consequence, we now have the foundation for the use of submillimeter wave spectroscopy in the identification and characterization of DNA macromolecules.

13.2.2. T-ray imaging

T-ray imaging uses the properties of absorption and transmission of different materials in the THz range. Information is usually presented in space and/or time domains. The imaging in this wavelength range has different applications in medicine and security. The entire THz range is strongly absorbed by all the polar liquids, like water. A picture of an object can be realized to determine the presence of water. In order to have an example of contrast and the spatially resolved measurement achieved, a T-ray image is taken of a vegetal leaf Fig. 13.2. We observe the contrast between location of water (in dark) and with a less moisture content (in grey). The main advantage of such a technique is the non-destructive action of the T-ray. As a consequence, the measurement can be repeated in the time domain as a video.

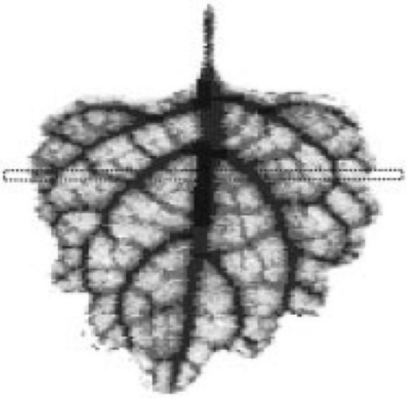


Fig. 13.2. Terahertz image of a leaf from a common houseplant, Coleus. The false color scale is correlated with water content, with darker part indicating more water. [Reprinted with permission from IEEE Journal of Selected Topics in Quantum Electronics, Vol. 2, No. 3, Daniel M. Mittleman, Rune H. Jacobsen, and Martin C. Nuss," T-Ray Imaging," fig. 5, p. 684. Copyright 1996, IEEE.]

THz wavelengths penetrate most of dry, non-metallic and non polar objects like plastics, paper cardboard and non polar organic substance. This property added to the capability of the range to determine the chemical composition of samples. It allows us to imagine a package inspection with a chemical content mapping, and the direct industrial application of this technology is food inspection. This inspection can be realized in real time with a digital signal processor to extract compositional information for the entire object. This wavelength range seems to be a completely open for medical imaging [Humphreys et al. 2002]. Both in dermatology and dentistry, this technique has yielded significant results. Two techniques are conceivable: transmission imaging of thin, clinically prepared tissue samples and reflection imaging. The first technique is practicable only *in vitro,* whereas the second one is possible *in vivo*. These two techniques are both valuable and depend on the task at hand. For dermatology purposes, detection of a cancer at the lowest layer of the epidermis can be achieved. T-ray imaging is able to give the size, the shape and the depth of the diseased tissue. The complete understanding of this detection method is still currently under development, and the chemical composition, or the different water absorptions are attributed to the change of absorption in the THz range [Humphreys et al. 2002].

Concerning dentistry, T-ray is able to provide two types of information. By absorption measurement dental caries, also known as tooth decay or cavities, are detected. Since caries are simply cavities or holes in the tooth, when a T-ray is sent in, the caries can be detected (Fig. 13.3). By measuring the difference of refractive index over the surface of a tooth, a map can be created. The presence of dentine and enamel, or only dentine, can be identified too by this measurement.

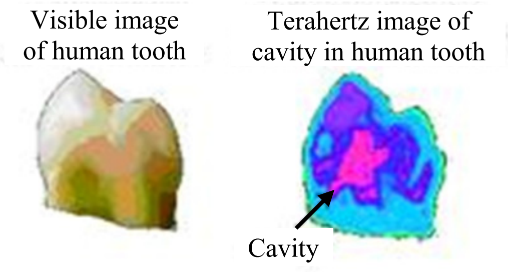


Fig. 13.3 Visible image of a human tooth and THz image of a tooth, on the second picture, a cavity is revealed. [Reproduced with permission from http://www.teraview.com/terahertz/id/35, Copyright 2009, TeraView Ltd.]

To finish with the T-ray imaging applications, metals are totally opaque to this wavelength range. As a consequence, weapons and knifes can be detected easily by this technique of imaging. This type of security system could replace the actual X-ray system in the airport [Clery et al. 2002]. Two main advantages are easily understandable. First, as T-ray is a non ionizing beam, this technique cannot damage any material. Secondly, a chemical spectrometer can be added to the security system in order to prevent any chemical or biological hazard.

13.2.3. THz research tool

THz spectroscopy can be used as a tool to investigate the structural property and carrier transport of nanostructures. The THz Time domain spectroscopy has been used in two particular nanostructures, first in photonic crystals and secondly in quantum cascade lasers. For the first structure studied, Prasad et al. reported a measurement of the normal incidence transmission coefficient of a photonic crystal slab unit with hexagonal arrays of air holes in silicon [Prasad et al. 2007]. For this experiment, THz radiation is generated and detected using photoconductive antennae. The wave propagates in a perpendicular direction to the plane of the photonic crystal slab and the transmitted electric field is measured as a function of time. The ratio of the Fourier transform of a pulse transmitted data through the sample to a reference one gives rise to the transmission spectrum.

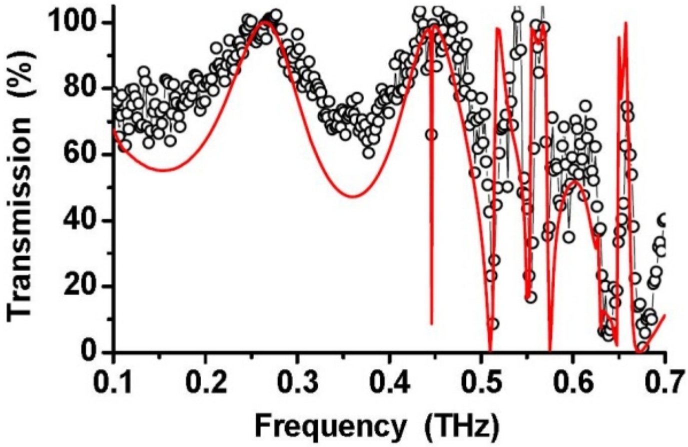


Fig. 13.4. Normal-incidence transmission spectrum of a photonic crystal slab with r = 150 µm, a = 400 µm, and t = 250 µm. The open black circles are experimental results, while the solid curve is obtained from simulations based on the finite elements method (FEM). [Reprinted with permission from Optics Express Vol.15, No.25, Tushar Prasad, Vicki L. Colvin, and Daniel M. Mittleman, "The effect of structural disorder on guided resonances in photonic crystal slabs studied with terahertz time-domain spectroscopy," fig. 3, p. 16958, Copyright 2007, Optical Society of America (OSA).]

Typical experimental results are presented in open black circles on Fig. 13.4. A comparison between the experimental results and a finite elements simulation method reveal the quality of the sample. The transmission spectrum is the superposition of two different responses. First at lower frequencies, Fabry Perrot oscillations are prominent and attributed to the finite slab thickness. Secondly, at higher frequencies, photonic crystal modes are excited. Transmission minima are the signatures of the guided resonances and the widths of its features represent the lifetime modes. Thanks to this measurement, the authors investigated the effects of several types of structural disorder in photonic crystals, and succeeded in determining the key parameters which lead to these modes, the disorder in the lattice periodicity.

The second structure studied by THz spectroscopy is the quantum cascade laser [Kröll et al. 2007]. A detailed understanding of laser action should help to improve the performances of these devices. Thus, the carrier dynamic behavior in such a complex systems is studied by time domain spectroscopy within the coherence time of the optical transition. THz pulses are generated using a femto-second laser Ti:Sapphire, exciting a photoconductive switch. The THz wave is transmitted through the THz QCL waveguide and detection is operated by a coherent electro-optic system, which senses the instantaneous electric field vector of the transmitted THz pulses. From the Fourier transform of the time dependent signal, two figures can be extracted, the absorption spectrum and the phase. The absorption measurement gives rise to a pronounced peak which represents the spectral gain of the THz QCL. The frequency behavior of the phase is the evidence for population inversion when the QCL is biased. In a second experiment using the time domain spectroscopy of a QCL, authors put in evidence the spectral gain variation of a bound to continuum THz quantum cascade laser as a function of current density [Jukam et al. 2008]. Thanks to these THz time domain spectroscopy measurements, a correlation between the spectral gain width and the alignment of the injector state with the active region has been determined as a function of the applied electric field. This spectroscopic tool has the principle advantage of being usable for both above and below threshold operations.

13.3. Broadband terahertz sources

THz sources can be categorized into broadband and narrowband. Two different approaches are possible to generate broadband THz radiation, electronically or optically.

The frequency multiplier is the first THz emitter we are presenting in this section. This device is composed of two main parts. The first part is a microwave oscillator emitting in the W-band (75–111 GHz). Then, a nonlinear solid state frequency multiplier, like a Schottky barrier diode is used. This kind of metal-semiconductor junction has the property of very short switching times on the order of 100 ps. As a consequence, broad bandwidth emission is operated in a range from 200 GHz to 1600 GHz with 30 mW and 1 μW respective output power [Siegel et al. 2002].

During the 60’s, a semiconductor device was realized which emits millimeter-wave frequencies, its name is IMPATT, which means IMPact Avalanche Transit Time. The acronym of this device directly comes from the physical mechanism which leads to the wave oscillation and amplification. These two mechanisms are due to the frequency dependent negative resistance. Avalanche breakdown and transit time effects are responsible for a phase difference between the current and voltage which give rise to this behavior of the negative resistance.

Terahertz tube sources, backward wave oscillators or so called Carcinotrons, are vacuum tubes based on emission from bunched electrons spiraling about in strong magnetic fields. The backward wave oscillators have one main advantage, they are continuously tunable electrically. An electron beam generated by an electron gun is moving in vacuum through a slow wave structure. Along the backward wave circuit, an *rf* wave is propagating. The longitudinal component of the slow wave electric field modulates the electron velocity. As a consequence an exchange of energy between the electron kinetic energy and the wave electromagnetic energy takes place. A magnetic field is applied in order to guide the electron beam along the circuit. The slow wave structures are periodic and allow filtering with pass bands and stop bands. These sources are able to generate microwaves up to 1200 GHz, with milliwatt level power, and operate between 2 and 6 kV and a magnetic field of 1 T [Siegel et al. 2002].

In another approach, using optical conversion, we present a technique based on femtosecond pulses generated by ultra short pulsed lasers, for example Argon laser pumped Ti: Sapphire laser. Emission is operated in an integrated circuit. Materials like silicon-on-sapphire are photoconductors, and the conductivity increases under illumination. The laser beam illuminates the photoconductor material with closely spaced electrodes which are sitting under a large DC voltage bias. As the excitation laser pulse is a femtosecond one, it generates an ultrafast dipole when the photoconductivity between the electrodes abruptly turns on and off at the location of the laser pulse. The time dependence of the electrical pulse created, and the transient electric dipole responsible for the THz emission are the same. Carriers generated by the laser beam are accelerated by the DC electric field between the two electrodes. The current created has frequency components that reflect a THz rate since the pulse duration has a timescale of the order of a picosceond. Coupled to a RF antenna, THz wave emission is produced. In order to increase the efficiency of this emitter, one can fabricate this source at the focal point of a spherical lens, allowing the radiated THz pulse to be collimated and focused on a detector [Fattinger et al. 1988].

We finish our discussion of broadband sources with a recently demonstrated way to generate THz radiation based on a gated two-dimensional electron gas. As predicted by Dyakonov et al. [1993] and demonstrated by Knap et al. [2004], the emission of electromagnetic radiation at plasma wave frequencies is generated because of instabilities in the steady state of the current flowing through a field effect transistor. The plasma generation which induces the THz emission is operated in a InGaAs/AlInAs high electron mobility transistor (HEMT). A threshold like behavior is observed. A certain value of current is required in order to obtain an increment of the plasma wave amplitude which exceeds the losses due to electron collisions with impurities and/or lattice vibrations.

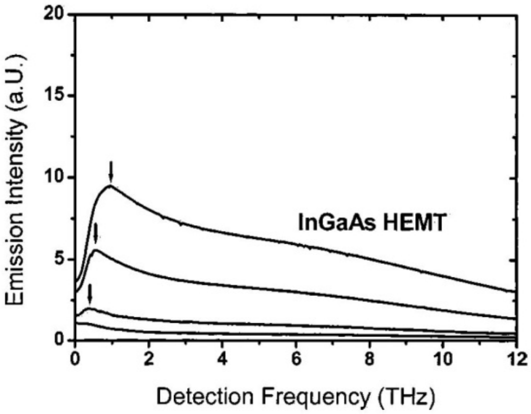


Fig. 13.5. Spectra of emission from an InGaAs HEMT for different source-drain voltages, U_{sd}. The arrows mark emission maxima at 0.42, 0.56, and 1.0 THz for U_{sd} equal to 0.3, 0.6, and 0.8 V, respectively. [Adapted with permission from Applied Physics Letters Vol.84, No.13, W. Knap, J. Lusakowski, T. Parenty, S. Bollaert, A. Cappy, V.V. Popov, and M.S. Shur, "Terahertz emission by plasma waves in 60 nm gate high electron mobility transistors," fig. 2, p. 2332, Copyright 2004, American Institute of Physics (AIP).]

Fig. 13.5 represents the spectral emission from an InGaAs HEMT for different source-drain voltages. The spectra present main peaks ranging from 0.42 to 1 THz. The current drain threshold (drain voltage) in this structure with 60 nm long gate is measured to ~4.5 mA (~200 mV). Estimation of power is in the nano-Watt range. Adjusting the design of the structure, a room temperature broadband THz emission peaked at 1.5–2 THz has been reported [Dyakonova et al. 2006].

13.4. Narrow band terahertz sources

13.4.1. Optical converter

A photomixer consists of two sets of interdigitated metal electrodes deposited on the surface of a low temperature grown material, for example GaAs. The photomixing approach allows the emission of radiation from microwave frequencies up to 3.8 THz upon exposure to two frequencies of laser light whose difference frequency gives an intensity modulation at the desired THz frequency [Brown et al. 1995]. Electron-hole pairs are generated by photon absorption pumped by two continuous waves in a low temperature grown GaAs. This material has remarkable photoconductive properties and sub-picosecond electron-hole recombination times. Photocurrents generated in the gaps between the electrodes are coupled to an antenna that emits the THz wave (Fig. 13.6). The authors report a continuous wave, and a coherent output in the range 30 to 1000 µm. The emitted light is in a narrow-band, and can be tuned over the full terahertz band by slightly shifting the optical frequency of one of the two lasers. Two 780–820 nm Ti: sapphire and 850 nm distributed Bragg reflector semiconductor lasers have been used. Typical output power of 1 W at 1 THz to 0.1 W at 3 THz is reported [Siegel et al. 2002].

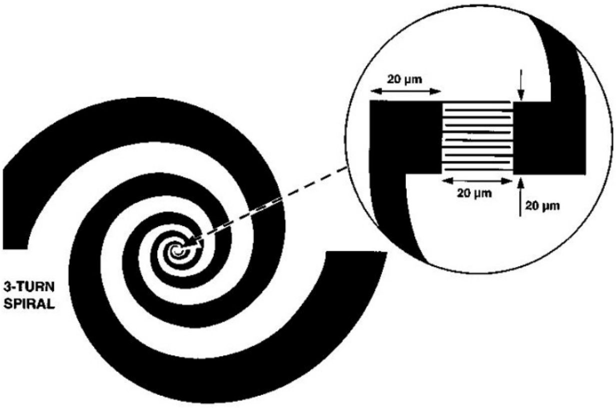


Fig. 13.6. Three-turn self-complementary spiral antenna. The inset shows an expanded view of the driving point of the antenna and the interdigitated electrode photomixer structure. [Reprinted with permission from Applied Physics Letters Vol.66, No.3, E.R. Brown, K.A. McIntosh, K.B. Nichols, and C.L. Dennis, "Photomixing up to 3.8 THz in low-temperature-grown GaAs," fig. 1, p. 285, Copyright 1995, American Institute of Physics (AIP).]

Another approach to generate THz radiation is to use an optical parametric oscillator (OPO). This radiation comes from a non linear effect in a crystal. The crystal is pumped by a laser source oscillating at a frequency ω_p, and two output waves of lower frequencies are emitted. These waves are called idlers with a frequency ω_i and source with a frequency ω_s. Obviously for energy conservation, the sum of output frequencies has to be equal to the input frequency ($\omega_i + \omega_s = \omega_p$). Using a yttrium-aluminum-garnet (YAG) laser as pump that generates pulses at λ = 1064 nm and a dielectric MgO:LiNbO$_3$ nonlinear crystals, Guo et al. [2006] succeeded to generate emission in a range from 0.6 to 2.4 THz with a narrow linewidth of 50 MHz. This technique is a field of active research for four main reasons. First, this experimental approach allows us to make compact components. Second, this source is widely tunable. Third, THz waves have a narrow-linewidth. And fourth, this source operates at room temperature. The experimental setup is composed of a pump source, a YAG laser that generates single longitudinal-mode pulses at a wavelength of 1064 nm, a seed source which is a continuous-wave external cavity laser diode. The wavelength can be tuned coarsely from 1056 to 1083 nm. As a final step, the gain media is made to consist of two cascade non-linear MgO:LiNbO$_3$ crystals. The THz-wave frequency is determined from the frequency difference between the pump and idler beams.

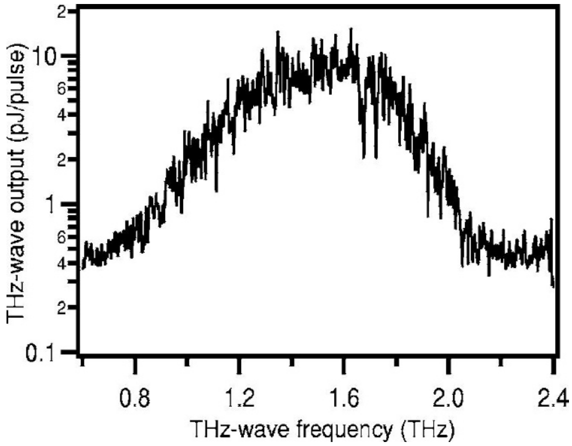


Fig. 13.7. THz-wave output energy in the tuning region of 0.6–2.4 THz with the pump input energy of 17 mJ/pulse and the seed beam power of 250 mW. [Reprinted with permission from Applied Physics Letters Vol.88, No.91120, R. Guo, K. Akiyama, H. Minamide, and H. Ito, "All-solid-state, narrow linewidth, wavelength-agile terahertz-wave generator," fig. 2, p. 91120-2, Copyright 2006, American Institute of Physics (AIP).]

Fig. 13.7 presents the THz wave output energy per pulse from 0.6 THz to 2.4 THz obtained by tuning the seed beam. This technique allows a wide spectral range of emission up to 7 THz using the difference frequency generation with a YAG laser and a non linear crystal of GaP [Tanabe et al. 2003].

The last optical converter we are presenting in this section is based on difference-frequency generation (DFG). This experimental approach to generate THz wave has some advantages over the OPO technique. First DFG offers relative compactness, tuning is more easily operated, and much lower pump intensities are required. The principle of difference frequency generation is based on an excitation of the phonon-polariton mode, under small-angle noncollinear phase matching conditions. The source used for the DFG is usually a YAG laser. As in the previous approach, the laser is separated into a pump and a signal beam, for example the pump beam is the YAG laser itself, and the third harmonic of this laser is sent in an optical parametric oscillator (β-BaB$_2$O$_4$) to generate the idler beam. The non-linear crystal was used by Shi et al. [2002] is GaSe. This crystal is reported to be particularly suitable for the THz range because its absorption coefficient is very low. The output THz wavelength can depend both on the idler wavelength and the external phase matching angular between the ordinary and extraordinary polarization of the beams inside the crystal. These latter authors succeeded in observing a continuous and coherent radiation in a range from 0.18 THz to 5.27 THz. With Cr:Forsterite laser media as the pump and idler, and GaP crystal, Suto et al. [2005] succeeded to emit in a range from 0.3 THz to 7.5 THz. A peak output power of 100 mW is obtained when the power of the two input beams was 3 mJ each. And to finish with, continuous-wave (CW) single-frequency THz waves were generated using difference-frequency generation in GaP. This optical alignment enabled to generate frequency-tunable CW THz waves in the 0.69–2.74 THz range [Nishizawa et al. 2006].

13.4.2. Optically pumped gas lasers

Inguscio et al. [1986] made an extensive review of all the optically pumped gas lasers in the far infrared range. In 1970, Chang and Bridges reported the first far infrared emission obtained by optical pumping. With methyl fluoride gas optically pumped by a CO_2 laser, six rotational transitions grouped in pairs near 452, 496, and 541 µm lead to a lasing emission. Population inversion is operated in some of the rotational levels of the excited vibrational state of gas molecules displaying a permanent electric dipole moment, thanks to an excitation from a CO_2 laser. Hundreds of gases exhibit FIR laser emission in the spectral region from 0.1 THz (2.9 mm) and 8.5 THz (35 µm). The radiative transition occurs between two molecular levels. As a consequence, on the one hand, an inhomogeneous broadening appears in the spectral ray due to the Doppler effect, and on the other hand, collisions and saturation result in a homogeneous broadening. Usually, the necessary pump power is high (20–100 W). The pump laser is injected into low-pressure flowing-gas cavities that emit in the terahertz range with an output power level of 1–20 mW [Siegel et al. 2002].

13.4.3. Semiconductor source based on Silicon and Germanium

THz stimulated emission has been achieved in silicon materials. Emission from shallow donor transitions in silicon is reported. Lasing with a wavelength of $\lambda = 59$ µm due to the neutral donor intracenter $2p_0 \rightarrow 1s(E)$ transition in Si:P pumped by CO_2 laser radiation is obtained [Pavlov et al. 2000]. The intracenter population inversion is based on the accumulation of charge carriers in the long-lived bound excited state of neutral donors. In the present case, phosphor belongs to the group V element, so this bound state in *Si* is usually a $2p_0$ state. As represented in Fig. 13.8, radiative transition occurs from $2p_0$ state to $1s(E,T)$ states in Si:P.

These levels are split off by the intervalley interaction. Lifetime calculations predict that the *2p$_0$* lifetime is 100 time longer than *1s(E)* at 5 K. CO$_2$ laser allows the optical excitation of free carriers into the conduction band. Then, these electrons lose their energy by optical and acoustic phonon emission and populate the upper laser state *2p$_0$*. This stimulated emission has been demonstrated also with another donor in silicon, antimony [Pavlov et al. 2000]. Si:Sb exhibit a stimulated emission at $\lambda = 58.2\ \mu\text{m}$ from the intracenter transition under CO$_2$ laser pumping. This process is limited in temperature, because above a lattice temperature of 30 K, intervalley and optical-phonon interactions with these centers are not negligible anymore.

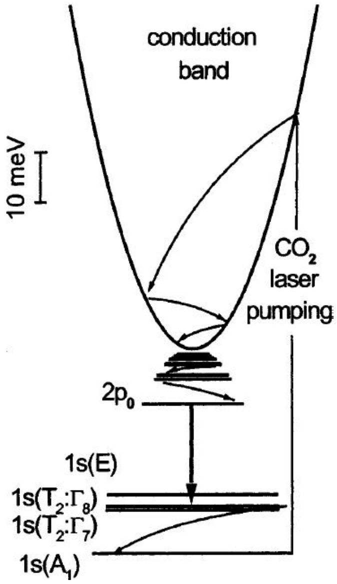


Fig. 13.8. Scheme for optical and non-radiative transitions in Si:Sb under CO$_2$ laser pumping: straight arrow down indicates the THz laser emission; curved arrows down indicate relaxation of the photoexcited electrons due to emission of acoustic and optical phonons. 1s(A$_1$) is the ground state of the Sb neutral donor, 1s(T$_2$), 1s(E) and 2p$_0$ are the excited states. [Reprinted with permission from Journal of Applied Physics Vol.92, No.10, S.G. Pavlov, H.W. Hubers, H. Riemann, R.K. Zhukavin, E.E. Orlova, and V.N. Shastin, "Terahertz optically pumped Si:Sb laser," fig. 1, p. 6633, Copyright 2002, American Institute of Physics (AIP).]

Germanium is another bulk material of group-IV that is used to emit coherent light in the THz region. This narrowband stimulated THz emission from the light hole state occurs in a crossed electric and magnetic field. The magnetic field is applied parallel to the $[110]$ and electric field parallel to the $[1\bar{1}0]$ crystallographic directions. Under magnetic field, subbands turn into Landau levels. A broadband emission is observed at low magnetic field

($1\ \mathrm{T} < B < 2\ \mathrm{T}$) due to a radiative transition from light to heavy hole states and at high magnetic field ($3\ \mathrm{T} < B < 4\ \mathrm{T}$), a narrowband stimulated emission is measured due to a cyclotron resonance emission, between two light hole Landau levels separated in energy by $\hbar\omega_c = \hbar eB/m^*$. The emission is tunable thanks to the magnetic field, and authors report an emission in a wavelength range from 1.9 THz (154 µm) to 2.5 THz (118 µm) [Unterrainer et al. 1990]. This experimental approach to achieve THz emission has been improved with Be-doped Ge crystals to reach a continuous excitation [Bründermann et al. 2000]. The limiting factor for the use of this type of source comes from the cooling system. These experiments are carried out at low temperature (T = 4.2 K), and an efficient cooling system is necessary due to the high input power (10–15 W).

Still using the same material, *p*-type Ge, but without the use of a magnetic field (in order to simplify the experimental set-up), laser emission has been observed. Instead of a magnetic field, a high pressure is applied to the crystal. This high pressure implies an uniaxial stress on the crystal and gives as a consequence, a splitting of the light and heavy hole bands and their respective acceptor states. A hole population inversion is achieved within these acceptor states with the excitation of an electric field under uniaxial stress. The laser emission is tunable by the applied uniaxial stress. For example, under a pressure of 7 kbar, emission was measured from 4.8 to 5.6 THz under pulsed mode [Altukhov et al. 1996]. Continuous wave laser action has been reported for these stressed lasers [Gousev et al. 1999].

In Si/SiGe/Si quantum well (QW) structures *p*-doped with boron, a THz emission has been reported [Kagan et al. 2000]. This emission is attributed to a similar mechanism as for *p*-doped Ge bulk material. The population inversion is achieved between strain-split off acceptor levels. Strain does not come from an external source. As for bulk materials, it is due to an internal strain resulting from the lattice mismatch in the multi-quantum well structure. When impurities are added to the SiGe/Si QW system, all relevant energy levels are affected by the electric field produced by the charge redistribution in the system. A strong electric field in the QW may favor the formation of resonant states for THz lasing. The pulsed electric field was applied along the SiGe layer, parallel to the interfaces. The measurements were performed at liquid He temperature. The holes in the ground states are impact-ionized into the valence band, and then electrically pumped by a strong external electric field into the resonant states. The transition from the resonant states to the ground states gives rise to the THz radiation. Emission was measured in a frequency range between 2.9 THz (103 µm) and 5.8 THz (51.7 µm) [Kagan et al. 2000].

For most of the coherent sources just presented, a pump laser is required for room temperature emission. With semiconductor sources, like *p*-doped germanium and Si/SiGe multi-quantum well structures, no pump laser is necessary, but both devices are operating at low temperature and for bulk germanium, a magnetic field or a high pressure are required. So, THz emission by a solid state device operating at room temperature has not yet been demonstrated. In the next section, we present a device emitting in the THz range, the quantum cascade laser (QCL) and extend the presentation to theoretical propositions to achieve a room temperature emission with such a structure.

13.5. Quantum cascade terahertz sources

13.5.1. GaAs based terahertz QCLs

As seen in previous sections, the emitters available in the THz range generally are not compact, transportable and easy adjustable in wavelength. The number of applications in this range of the electromagnetic spectrum are so important, that succeeding to achieve THz emission with such a device would be of crucial interest, especially if room temperature operation is achieved.

The discussion about the quantum cascade lasers in Chapter 7 is concentrated on the mid-infrared portion of the electromagnetic spectrum. In fact, the idea of QCL can be applied to the THz portion as well. The first working THz QCL was reported in 2002 by Köhler et al. [Khöler et al. 2003]. The authors report an emission at 4.4 THz up to 50 K, with an output power up to 2 mW. Two challenges were overcome by this observation. First, as the radiative energy transition is very small compared to the mid-infrared one, the carrier injection has to be very selectively tuned towards the upper laser state, and the extraction has to be very efficient to reach population inversion. This condition has been achieved thanks to the design of a “chirped superlattice” active region. The structure is composed of seven wells made with GaAs and barriers made with $Al_{0.15}Ga_{0.85}As$. Secondly, in the THz range, the overlap between the active region and the optical mode is very small due to the long wavelength, and the maximum thickness one is able to grow by MBE. So, this group suggested to put the active region between two heavily doped layers in order to confine the mode. The question of waveguide design has extensively been described in two recent reviews [Scalari et al. 2008], [Williams et al. 2007].

In order to improve working conditions like temperature or continuous wave, different structure designs were proposed. The first structure is called the chirped superlattice, which is similar to the mid-infrared QCL [Colombelli et al. 2001] [Rochat et al. 2002]. Laser action originates from inter-miniband transitions in active regions. Secondly, there is another design called the bound-to-continuum that has also been suggested [Scalari et al. 2003a]. In this structure, injection is more selective because special attention is paid to the injector. The injector is isolated in a minigap. This structure is the first to exceed a working temperature above liquid Nitrogen, Indeed the highest temperature reached is 90 K. Thirdly, a design named bound-to-bound, in this structure electrons are injected into the upper laser state from a discrete injector level, and extraction of carriers is operated by LO-phonon emission. This mechanism is very efficient, and allows a rapid transfer of electrons from one cascade to the next one. The first design is operating at 3.4 THz in a four wells active region up to 69 K in pulsed mode [Williams et al. 2003]. The efforts were mostly concentrated on this type of structure, and the best performance to date was achieved without a magnetic field, with a design using this depopulation process in a three-well structure, with an operating temperature of 186 K in pulsed mode [Kumar et al. 2009].

Magnetic field has extensively been used as a powerful spectroscopic tool in QCL, as well in mid-infrared and in THz ranges. Thanks to the quantization of the carrier levels in the three directions, the scattering processes that limit the optical emission could be identified in these devices as the one or two electron scattering processes [Kempa et al. 2002], and the LO-phonon emission by hot electrons [Péré-Laperne et al. 2007]. Laser action has also been observed only under strong magnetic fields for some special designs called “big well" [Scalari et al. 2003b] [Scalari et al. 2004]. Thanks to the magnetic field, in the extreme quantum limit, no scattering process is able to depopulate the upper laser state, and as a consequence, the lifetime increases and the population inversion is easier to achieve. Recently, the highest working temperature has been achieved thanks to the use of a magnet. Laser operation is reported at a frequency of 3 THz up to 225 K [Wade et al. 2008]. The structure is a “bound to bound” design. Owing to the suppression of all non-radiative scattering processes, laser action is achieved at a much higher temperature.

13.5.2. InP based terahertz QCLs

GaAs based material system is suitable for THz QCLs because one is able to easily change the band offset by changing the Aluminum content without any lattice mismatch. One attempt to make an InP based growth for a THz QCL has been realized, the structure is composed of $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ [Ajili et al. 2005]. The active region is based on a bound to continuum transition, and the emission occurs at 3.6 THz. The main incentive to use this material compared to GaAs/AlGaAs is the value of the effective mass in the well. In $In_{0.53}Ga_{0.47}As$, $m^* = 0.043m_0$, whereas in GaAs, $m^* = 0.067m_0$. This property has three main advantages, first a low effective mass means a large oscillator strength, secondly in this material, the thicknesses are more important, so it is expected to decrease the interface roughness effect, and thirdly, the LO phonon scattering rate decreases as the effective mass decreases. The waveguide exploits a combination of metallic and dielectric confinement. Laser action has been measured for temperatures ranging from 10 up to 45 K. The collected peak power at 10 K reaches a value of 250 µW and a threshold current density of J_{th}=460 A/cm^2. The performances of this InP based QCL are lower than GaAs based structures. Vasanelli et al., studied the laser in a magnetic field using a mid-infrared InP based structure. They suggest that alloy disorder in the well can be a dominant elastic scattering process and claim that this is the mechanism which limits the laser performances [Vasanelli et al. 2006].

A summary for the best reported THz QCLs of different designs is shown in Table 13.1.

	Material	Energy, Frequency	J_{th} (A/cm^2)	Max T (pulsed)	Max power (pulsed)	Max T (CW)	Max power (CW)
Chirped superlattice	GaAs/ AlGaAs	18 meV, 4.4 THz	165–185 (pulsed/ continuous)	67 K	90 µW (67 K) 4.5 mW (10 K)	45 K	500 µW (40 K) 2.5 mW (10 K)
Bound to continuum	GaAs/ AlGaAs	12 meV, 2.9 THz	105	95 K	15 mW	70 K	15 mW
Bound to bound with LO phonon depopulation	GaAs/ AlGaAs	16.1 meV, 3.9 THz	830–410 (pulsed/ continuous)	186 K	Few µW		
Bound to continuum with LO phonon depopulation	GaAs/ AlGaAs	14.6 meV, 3.7 THz	178	116 K	4 mW (80 K) 10 mW (30 K)	53 K	5 mW (40 K)
Bound to continuum with LO phonon depopulation	InGaAs/ InAlAs on InP	14.9 meV, 3.6 THz	460	45 K	250 µW (10 K)		

Table 13.1. Best reported performances of the different THz QCL designs.

13.6. Magnetic field effects

A magnetic field is a very useful experimental tool to study the two-dimensional systems. When a magnetic field is applied perpendicular to the heterostructure layers (parallel to the growth axis, usually labeled z), the 2D system confined in one direction of space (z) with carriers free in the plane (x, y) becomes confined in the three directions of space (x, y, z). For a magnetic field perpendicular to the layers, electrons are not affected in the direction z but they describe a cyclotron orbit in the plane (x,y). When the radius of the cyclotron orbit is comparable to the de Broglie wavelength, a quantum treatment gives rise to a discrete energy spectrum of Landau levels.

Thus, a high magnetic field applied to a quasi-two-dimensional structure converts the system into a quasi-0D structure [Ferreira et al. 1991]. The energy of the eigenstates, neglecting the spin splitting in this completely confined system is given by:

$$\text{Eq. (13.1)} \qquad E = E_m + \left(n + \frac{1}{2}\right)\hbar\omega_c$$

Where, E_m is the subband energy, n the Landau level index and ω_c is the cyclotron frequency given by $\omega_c = e\,B\,/\,m^*$.

This experimental technique has been used with different QCLs. In the mid-infrared range, very high magnetic fields are necessary, most of the experimental results were obtained in the high magnetic field facilities, like the one at Laboratoire National des Champs Magnétiques Pulsés, in Toulouse, France or the National High Magnetic Field Laboratory at Florida State University.

For two mid-infrared GaAs/AlGaAs lasers emitting at 11 µm and 9 µm, the magneto-transport and light intensity as a function a magnetic field have been measured [Smirnov et al. 2002a][Smirnov et al. 2002b]. In both cases, strong oscillations are reported for the current flowing through the structure as a function of magnetic field at a constant bias, and also for the laser power as a function of magnetic field at a constant current. These oscillations are attributed to the intersubband magneto-phonon resonance due to the electronic LO phonon scattering. At specific magnetic field values, the Landau levels of the two states involved in the laser transition are resonant with an emission of one LO phonon. As a consequence, electrons can scatter non radiatively from the upper laser state to the lower state. This non radiative scattering process increases the current flowing through the structure and decreases the emitted laser power. Magnetic fields help to reveal this scattering process by giving rise to some features at specific magnetic field values, and the magnetic field helps us to understand the limiting scattering process for a particular structure. Calculation of the electron-LO phonon scattering rates between Landau levels in these structures have been reported [Becker et al. 2004]. Becker et al. report strong oscillations of the upper laser state scattering rate as a function of magnetic field. These oscillations are directly linked to the current and light power oscillations, when the upper laser state lifetime increases, the light power increases, and the current flowing through the structure decreases. Another scattering process has been found in a GaAs/AlGaAs device as a result of magneto-transport and magneto-optic measurements. This scattering process is an elastic one, the scattering is due to the interface roughness [Leuliet et al. 2006]. In this structure, the authors give evidence of an elastic process which is dominant under the magnetic field, and with the help of calculations, the interface roughness scattering is found to be the dominant of these processes. All the following LO phonons, interface roughness, LA phonons, impurities and alloy scattering processes are compared at zero magnetic field. Interface roughness is comparable to LO-phonon emission in this particular structure. The sum of these processes accounts for all the features in the current and light power as a function of magnetic field.

In the far-infrared range, magnetic field studies are even more interesting than in the mid-infrared. Suppose that the energy difference between the laser states is about 20 meV (4.8 THz). In GaAs quantum well structures, the resonance between the fundamental Landau level of the upper laser state and the first excited Landau level of the lower laser state occurs at $B = 11.7$ T. This magnetic field range is achievable with a standard superconducting magnet. Therefore, it is possible to reach a high enough magnetic field which takes the system to be "over the extreme quantum limit". The "extreme quantum limit" corresponds to the magnetic field values above which all the excited Landau levels of the lower laser state are above the fundamental Landau level of the upper laser state. In this particular configuration, elastic scattering processes cannot scatter carriers from the upper state to the lower one, and as a consequence, the upper laser lifetime increases. For some designs, a very low threshold current density, and a high working temperature were reported.

A THz quantum cascade structure has been studied with the help of a magnetic field [Blaser 2002]. Radiative transitions occur in a big well. Two materials were studied InGaAs/InAlAs and GaAs/AlGaAs both grown by MBE. Since the materials are different for the two structures, band offsets and effective masses are also different. Electroluminescence measurements give the radiative transitions at energies 15 meV and 16 meV respectively. On both structures, Blaser et al. measure a strong oscillating behavior of the emitted light as a function of magnetic field. As in the mid-infrared range, the oscillations can be explained by scattering processes which modify the scattering rates of the levels which are involved in the radiative transitions. Because the cyclotron energy is a function of the inverse of the effective mass, the resonances, and also the maximum and minimum energies of the emitted light, are not the same for the two structures. The output power increases by a factor 6 as the magnetic field increases. This enhancement of emitted power comes from the crossing into the extreme quantum limit. We are going to discuss that further later. The first lasing operation measurement in a QCL under strong magnetic field was done in a GaAs/AlGaAs bound to continuum design emitting at 4.4 THz [Alton et al. 2003]. The authors report a drastic reduction of the threshold current density as a function of magnetic field from 290 A/cm^2 at 0 T to 160 A/cm^2 at 4.3 T.

Again, oscillations of current density and laser power are attributed to a modulation of the lifetime of the upper state of the laser transition. This is due to the combined effects of inter-Landau level resonances and a progressive quenching of non-radiative relaxation channels. The authors specify that the threshold current density is a function of both the upper and the lower laser lifetimes. The extraction efficiency is deduced to be a crucial key parameter in these THz structures as compared to the mid-infrared ones.

Three scattering processes are important in the THz QCLs. First, Kempa et al. presented a two electrons scattering process, or the so-called electron-electron scattering mechanism [Kempa et al. 2002]. This particular inelastic scattering process has its own signature in current density measurements as a function of magnetic field. Starting with the initial two electrons in the upper subband (Landau level index: *k, m*), and ending after the scattering with two electrons in the lower subband (Landau level index: *l, p*), the energy conservation rule leads to:

Eq. (13.2) $$\Delta = \frac{n}{2}\hbar\omega_c$$

where Δ is the intersubband energy separation and $n = p + l - k - m$. When the above condition is fulfilled, the electronic scattering rate of the upper state increases.

Secondly, it was discovered that a one electron scattering process in GaAs/AlGaAs QCLs, plays a major role, and that is the interface roughness [Kempa et al. 2002] [Péré-Laperne et al. 2007]. Parameters which specify the roughness at the interface are the height of the roughness Δ and the correlation length Λ. The authors used these two parameters to describe the laser power oscillations. The typical value of the roughness height is half a monolayer: $\Delta = 1.5$Å and the correlation length: $\Lambda = 60$ Å. On the same sample, another scattering process was shown to be present, the emission of LO phonons. Even if the laser transition is lower than the energy of one LO phonon in GaAs, LO phonon emission is assumed to be responsible for the decrease of laser power at certain magnetic fields. In that particular structure, the hot electron population in the upper laser state gives rise to a series of oscillations in a magnetic field which is attributed to an inelastic scattering process. This process is the emission of LO phonons toward the lower laser state. With the help of these examples, we have been able to show how a magnetic field can be a powerful tool to understand the limiting factors of THz structures.

A magnetic field has also been used to increase the working range of THz QCL, and some structures were specially designed to reach population inversion under strong magnetic fields [Scalari et al. 2003b] [Scalari et al.

2004]. First, a structure based on a vertical transition in a "big well" only allows emission in a magnetic field. Three bound states are present in the quantum well. Population inversion is achieved between the two first excited states separated by 7.9 meV (1.9 THz) and 15 meV (3.6 THz), the depopulation of the lower state is achieved through the fundamental bound state of the well. Carriers are traveling through the structure to the next active region thanks to a miniband made out of five wells. As in previous studies, oscillations of laser power, current density and threshold current density are observed. In these devices, the oscillations are attributed to the modulation of the lifetime of the upper state of the laser transition as a function of magnetic field. The lifetime is a function of the Landau level resonances, and the progressive quenching of non-radiative relaxation channels. These structures are not lasing at zero magnetic field because population inversion cannot be achieved. But in a magnetic field, the upper laser state lifetime increases, and for some magnetic fields the population inversion condition is realized. When all the scattering processes are quenched, the extreme quantum limit is achieved, and laser action is observed. In this situation, the threshold current density is reduced and is measured to be 1 A/cm^2 at 13 T.

The highest lasing temperature, in THz QCL, has been recorded with the help of an applied magnetic field [Wade et al. 2008]. This result has been obtained with a bound to bound QCL with LO phonon depopulation process. The structure is composed of GaAs/Al$_{0.15}$Ga$_{0.85}$As, with 178 periods, and laser action is supposed to occur at 13 meV (3.1 THz) at zero magnetic field (Fig. 13.9a). With a fixed value of bias applied to the structure, and at a fixed magnetic field value (19 T), the highest working temperature of this QCL is increased from 178 K to 225 K. Two reasons explain this performance increase. First, for such values of the magnetic field, all the excited Landau levels of the lower laser state are above the fundamental Landau level of the upper laser state, so the system is in the extreme quantum limit. No elastic or inelastic scattering process can depopulate the fundamental Landau level of the upper laser state, and as a consequence, the lifetime of this level increases. Secondly, the depopulation of the lower laser state is increased, thanks to the magnetic field. At zero magnetic field, the lower laser state is depopulated by emission of LO phonons. At 19 T, an elastic scattering process is used to depopulate the lower laser state too. Indeed the first excited Landau level of the stage at which LO phonon extraction occurs is resonant with the fundamental Landau level of the lower laser state (Fig. 13.9b). In this way, the lifetime of the lower laser state is decreased. As a result of an increase of the upper laser state, and a decrease of the lower laser state, the working temperature of this QCL is increased about 45 K.

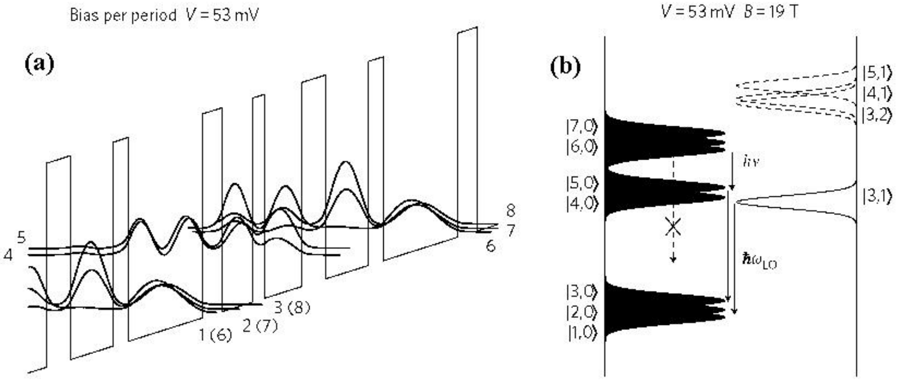


Fig. 13.9. (a) Conduction band diagram and squared wavefunctions of the relevant sub-band states. Levels 3,2,1 are levels 8,7,6 for the next period. At a bias of 53 mV per period, zero B-field, the laser transition takes place from 6 → 5 with fast LO-phonon assisted relaxation (5,4 → 3,2). (b) Illustration of the field-induced resonant enhancement of the lasing transition between LLs $|6,0\rangle$ *and* $|5,0\rangle$*. Inelastic scattering by LO-phonon emission from the upper laser transition state is quenched (dashed arrow), while elastic scattering from the final state* $|4,0\rangle$ *to another LL (*$|3,n\rangle$*) is allowed. [Adapted with permission from Nature Photonics Vol.417, p. 157, fig. 1a, Kohler, R, Tredicucci, A., Beltram, F., Beere, H., Linfield, E., Davies, A.G., Ritchie, D., Lotti, R., Rossi, F, "Terahertz semiconductor-heterostructure laser, Copyright 2008, Macmillan Publishers Limited.]*

We close the discussion concerning magnetic field effects on quantum cascade structures with reports from two groups who have made two color QCLs as a function of applied electric and magnetic fields [Wade et al. 2008] [Scalari et al. 2006]. The two groups used a different type of structure, firstly a bound to bound design [Wade et al. 2008], and secondly a big well structure [Scalari et al. 2006]. In both cases, by applying a different electric field on the structure, one can change the alignment of the upper and lower subbands. With the help of the magnetic field, and its effect on modifying the upper and lower state lifetimes, population inversion can be achieved for both values of the applied bias. The most amazing result is a two color (1 THz & 3 THz) laser emission for the same applied bias and the same magnetic field [Wade et al. 2008]. Wade et al. report this measurement at $B = 20.5$ T and a bias of 61 mV/period. To achieve these emissions, as before, the QCL has to be in the extreme quantum limit for the two upper laser states, and the depopulation of the second lower laser state has to be purposely operated by the LO phonon stage. The restriction to elastic scattering is made possible due to the presence of the magnetic field.

13.7. Difference frequency generation

Very recently, a room temperature THz emission in a mid-infrared quantum cascade structure has been reported [Pflüg et al. 2008] [Belkin et al. 2008]. The device is based on intra-cavity difference-frequency generation (DFG) in dual-wavelength mid-infrared QCLs with a giant optical nonlinearity monolithically integrated in the active region. This approach is a good way to avoid the limiting factor induced by THz QCLs. DFG is a nonlinear optical process involving two beams at frequencies ω_1 and ω_2. In a medium with second order nonlinear susceptibility $\chi^{(2)}$, these two waves interact and produce radiation at frequency $\omega = \omega_1 - \omega_2$. Due to the free-carrier absorption at THz frequencies, most of the light generated via intra-cavity DFG is lost. Also, the THz DFG output is expected to be strongly divergent. In order to overcome this problem, the authors used two approaches. First, they attached a silicon hyper-hemispherical lens on the device facet [Belkin et al.2008]. Secondly, they created a waveguide for a surface emission instead of edge emission [Pflüg et al. 2008]. The active region is based, as is the most efficient mid-infrared QCL, on InGaAs/InAlAs, lattice matched to InP material. The mid-infrared QCL is operated in pulse mode. The device emits at two mid-infrared wavelengths $\lambda_1 = 8.9$ µm and $\lambda_2 = 10.5$ µm and produces terahertz output at $\lambda = 60$ µm via difference-frequency generation. The band diagram and moduli squared of the wavefunctions are presented in Fig. 13.10a. The first period of the structure emits at λ_1 and the second period at λ_2. Fig. 13.10b puts in evidence the DFG principle and the terahertz energy difference. The output power achieved with the silicon lens are 7 µW at 80 K, 1 µW at 250 K, and 300 nW at room temperature. With surface emitting waveguide design, 1 µW output power at 80 K and 70 nW output at 300 K are reported.

These structures are the first successful devices to reach room temperature working conditions for a THz QCL. Owing to the mid-infrared capabilities, with watt level continuous wave power at room temperature, this device based on DFG may reach a microwatt level output power under these conditions. Improvement in the waveguide design, should in particular allow such power.

Another approach for room temperature THz QCL emission is presented in the next section. This approach is closer to the actual THz designs, but this time using a wide bandgap material.

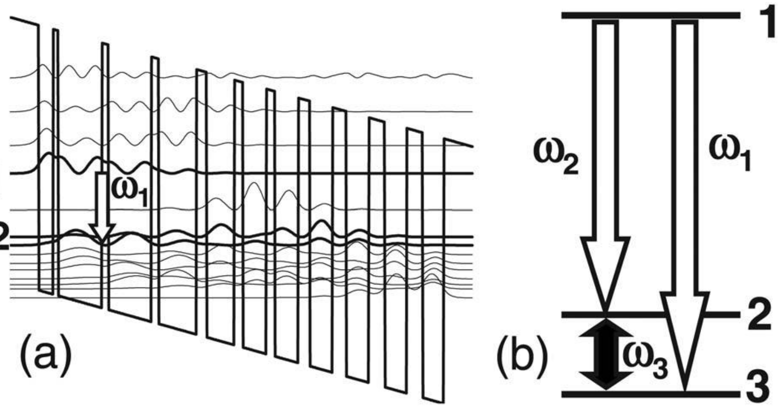


Fig. 13.10. (a) Calculated conduction-band diagram of one period of a bound to continuum quantum cascade laser section at applied bias of 37 kV/cm. The wavy curves represent the moduli squared of the wavefunctions. The electron states important for DFG are shown in bold and labeled 1 to 3. The layer sequence (in nm), starting from the injection barrier, for this structure is ***4.0****/ 2.4/* ***0.7****/ 6.5/* ***0.8****/ 6.4/* ***0.8****/ 5.8/* ***2.2****/ 4.0/* ***1.3****/ 3.8/* ***1.4****/ 3.7/* ***1.5****/ 3.6/* ***1.9****/ 3.6/* ***2.5****/ 3.6/* ***2.5****/ 3.5. (The layer sequence for one period of a double phonon resonance section which generates frequency* ω_2 *is* ***4.0****/2.0/* ***0.7****/ 6.0/* ***0.9****/ 5.9/* ***1.0****/ 5.2/* ***1.4****/ 3.8/* ***1.2****/ 3.2/* ***1.2****/ 3.2/* ***1.6****/ 3.1/* ***1.9****/ 3.1/* ***2.2****/ 3.0/* ***2.2****/ 2.9). The barriers are indicated in bold face and the underlined layers are doped to* $n=3\times10^{17}$ *cm*$^{-3}$*. (b) Diagram showing the DFG process between the electron states in the bound-to-continuum section. [Reprinted with permission from Applied Physics Letters Vol.92, No.201101, M.A. Belkin, F. Capasso, F. Xie, A. Belyanin, M. Fischer, A. Wittmann, and J. Faist, "Room temperature terahertz quantum cascade laser source based on intracavity difference-frequency generation," fig. 1, p. 201101-2, Copyright 2008, American Institute of Physics (AIP).]*

13.8. GaN QCLs for high temperature operation

There are two main issues in narrow band gap technologies that limit high temperature operation with demonstrated THz QCLs, both of which are related to optical phonons. The optical phonon energy, $\hbar\omega_{LO}$, in narrow band gap semiconductors is typically 30–40 meV. Fig. 13.11a presents a schematic view of a three subbands QCL. The radiative transition is supposed to occur from subband S_2 to subband S_1. Extraction of carriers is achieved by the most efficient process reported in QCL, emission of LO-phonon, from subband S_1 to subband S_0.

The first limitation is the thermal excitation of electrons from the ground state (S_0) into the lower laser state (S_1), which reduces the population inversion. For an electronic temperature of 300 K, 1/3 to 1/4 of electrons in the ground state (S_0) may be found in S_1 due to thermal activation. Moreover, the electron temperature has been found to be much higher than the lattice one. Microprobe band-to-band photoluminescence measurements, on a LO-phonon depopulation QCL, give a difference between the electron temperature of the upper laser state and the lattice temperature of about 100 K [Vitiello et al. 2005]. This experimental approach also allows to study another design, the bound to continuum. In such devices operating in continuous mode, the electron temperature in both the active region and the injector are found to be thermalized, and all the subbands share the same electron temperatures. The electron temperature is found to be higher than the lattice one in this design structure, too.

The second limitation is based on the non-radiative electron relaxation (leakage current). At low temperatures, as long as $S_2 - S_1 < \hbar\omega_{LO}$, the electron lifetime at S_2 is sufficiently long to have a good population inversion. At higher temperatures, however, electrons with higher kinetic energy on S_2 have enough total energy to emit an optical phonon and relax directly to S_1. This non-radiative current significantly decreases the optical gain. This scattering process, LO-phonon emission by hot electrons, has been demonstrated thanks to experiments with magnetic fields [Péré-Laperne et al. 2007]. The structure studied is one where the extraction process is mixed, involving a continuum of states, and a LO-phonon depopulation stage. As a result of these measurements, the electron temperature is estimated to be 150 K over the lattice one.

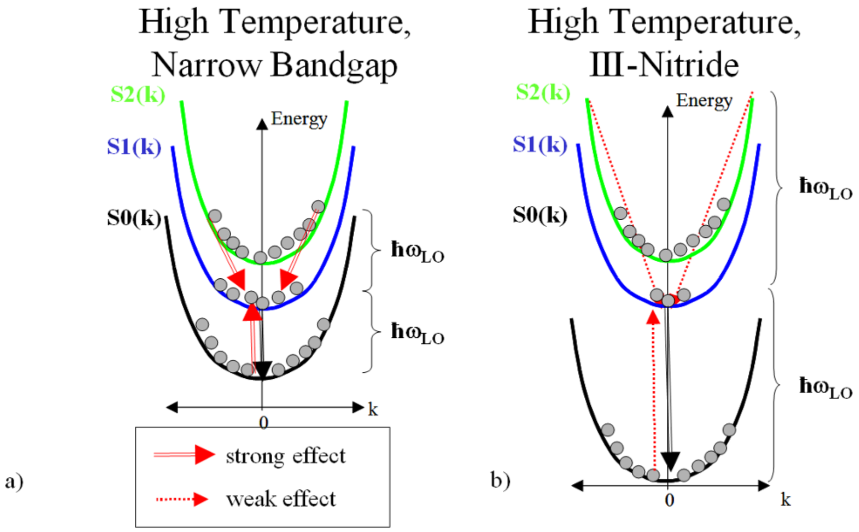


Fig. 13.11. High temperature processes in inter-subband THz emitter. a) Narrow bandgap material schematic. b) III-Nitride based emitter with larger optical phonon energy.

In order to overcome these disadvantages, III-Nitride-based systems have been proposed to exploit the large optical phonon energy (~90 meV) [Sun et al. 2005]. As shown in Fig. 13.11b, large bandgap materials are better for operation at higher temperatures. In this system, the energy separation between S_0 and S_1 can be increased significantly to keep the thermal population of S_1 low. Moreover, ultrafast LO phonon scattering (≤150 fs) has been observed in GaN/AlGaN QWs, which may make the population of S_1 even lower due to the fast relaxation of electrons from S_1 to the ground state. In addition, the large LO-phonon energy can also increase the lifetime of the upper laser state by reducing the relaxation of electrons with higher in-plane kinetic energy [Iizuka et al. 2000].

Recently, several groups have presented experimental evidence of intersubband transitions in GaN/AlGaN quantum wells [Gmachl et al. 2000] [Kishino 2000] [Hofstetter et al. 2003] [Hamazaki et al. 2004] [Baumann et al. 2005] [Friel et al. 2005]. These GaN/AlN heterostructures have very large conduction band discontinuities ($\Delta E = 1.75$ eV), which can accommodate inter-subband transitions at wavelength ranging as low as 1.1 µm, in the near infrared range. Molecular beam epitaxy has been used to realize AlN/GaN superlattices (SL) absorbing at optical communications wavelengths [Gmachl et al. 2000]. However, there are few reports for metalorganic chemical vapor-phase epitaxy (MOCVD)-grown AlN/GaN SL absorbing in near-infrared [Waki et al. 2003] and fewer leading to absorption around 1.55 µm [Baumann et al. 2006] [Nicolay et al. 2007]. Today, most nitride based commercial optoelectronic devices are grown by MOCVD. Thus, there is a lot of interest to realize high quality ISB absorption by MOCVD. AlN/GaN SL, can, despite the lattice mismatch of 2.4%, be realized crack-free and be grown pseudomorphically [Bykhovski et al. 1997]. However III-Nitrides are piezoelectric materials. In conventional c-plane growth these highly strained layers generate multi-MV/cm electric fields [Suzuki et al. 2002]. Thus, the interface or SL thickness fluctuations degrade the absorption quality significantly. Another problem is AlN/GaN or GaN/AlN interface stability, and their dependence on GaN or AlN template.

Room-temperature inter-subband luminescence from MBE grown GaN/AlN QWs has been already demonstrated at ~2 µm [Julien et al. 2007] [Nevou et al. 2007]. The structure is a SL composed of about 200 periods, and it is designed to exhibit three bound states in the well. Carriers are pumped from the lowest subband (first bound state) to the highest one (third bound state) and emission occurs from the third state to the second one. The quantum efficiency of this structure is very low, estimated to be 10 pW/W, but this demonstration is a promising step toward quantum cascade emitters.

Obviously, the high conduction band offset of GaN/AlN cannot be fully used for THz structure design, but using a ternary alloy ($Al_xGa_{1-x}N$) for the barrier instead of a binary AlN, the band offset can be decreased. For example, with 10% aluminum content, the conduction band discontinuity is $\Delta E = 175$ meV. An optically pumped structure has been proposed with barriers made of $Al_{0.07}Ga_{0.83}N$ and wells made of GaN [Vukmirović et al. 2005]. This design is made for an emission at 34 μm (8.8 THz) which is approximately an energy of 36 meV. This energy can never be obtained with AlGaAs/GaAs material as it is close to the LO phonon energy ($\hbar\omega_{LO}$). Fig. 13.12 presents the structure based on a two quantum wells design with three levels. Electron from the fundamental subband (level 1) are pumped to the upper subband (level 3), emission occurs in the THz range from upper subband (level 3) to an intermediary subband located 36 meV below. Next, carriers are extracted by LO phonon emission. The authors investigate the possible population inversion, and present a calculation of the output power. Room temperature operation is predicted with such a structure.

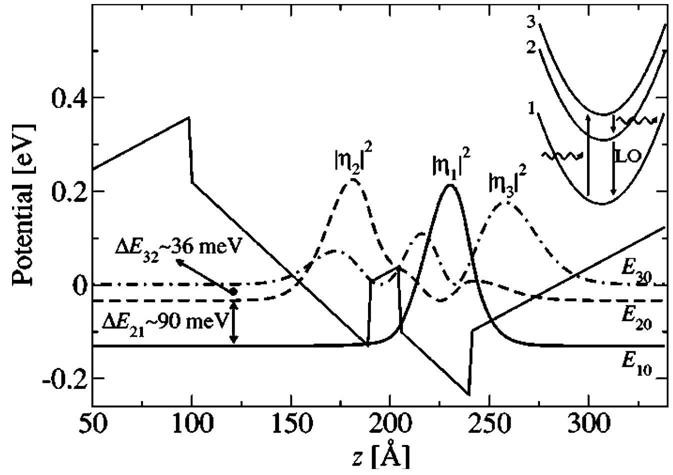


Fig. 13.12. Conduction-band profile, energy levels, and wave-function moduli for the structure x=0.070, L_{w1}=90 Å, L_b=15 Å, L_{w2}=35 Å. The inset shows the three-level scheme of this laser: electrons from the ground state (subband 1) are optically pumped to the upper laser level (subband 3), while a fast depopulation of the lower laser level (subband 2) is achieved via resonant LO-phonon emission. [Reprinted with permission from Journal of Applied Physics Vol.97, No.103106, N. Vukmirovic, V.D. Jovanovic, D. Indjin, Z. Ikonic, P. Harrison, and V. Milanovic, "Optically pumped terahertz laser based on intersubband transitions in a GaN/AlGaN double quantum well," fig. 1, p. 103106-1, Copyright 2005, American Institute of Physics (AIP).]

Other electrically pumped designs were also proposed. First in 2004, on two different growth planes (a-plane and c-plane), it was suggested to grow a four wells quantum cascade structure composed with 25% and 20% Aluminum content using $Al_xGa_{1-x}N$/GaN [Jovanović et al. 2004]. The difference between the two growth planes is that whereas on the a-plane, only the spontaneous electric field is taken into account, on the c-plane, both spontaneous and piezoelectric fields contribute to the conduction band profile.

For both structures, the injector is labeled 3’, and the carriers are injected into the upper laser state (level 2). Radiative transition occurs between subband 2 and 1 and then electrons are extracted from level 1 to level 0 by emission of one LO phonon (Fig. 13.13). For both structures, a self-consistent rate equation system is derived. The authors take into account the fast scattering processes like emission of LO-phonons, consider the electron-electron scattering mechanisms, and model the temperature dependence of the population inversion. They conclude that room temperature performance of such structures is feasible.

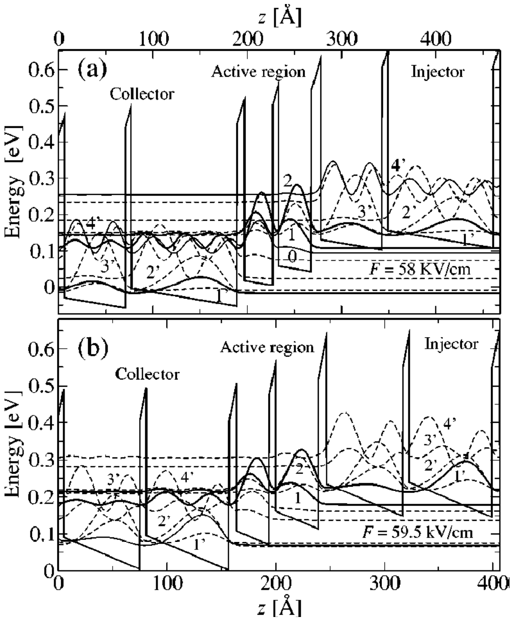


Fig. 13.13. A schematic diagram of the quasi-bound energy levels and wave functions squared for an injector-active region-collector segment of (a) a-plane and (b) c-plane GaN/Al$_x$Ga$_{1-x}$N (x =25% and 20%) QCL. The layer sequence of one period of the structures, in nanometers, from right to left, starting with the injection barrier, is (a) ***0.8****, 3,* ***0.6****, 3.5,* ***1****, 6.5,* ***0.6****, 11.2 and (b)* ***0.7****, 3,* ***0.6****, 3.9,* ***0.8****, 7,* ***0.6****, 7.6. The normal print denotes the wells and bold the barriers. [Reprinted with permission from Applied Physics Letters Vol.84, No.16, V.D. Jovanovic, D. Indjin, Z. Ikonic, and P. Harrison, "Simulation and design of GaN/AlGaN far-infrared ($\lambda \sim 34\ \mu m$) quantum-cascade laser," fig. 1, p. 2995, Copyright 2004, American Institute of Physics (AIP).]*

To conclude the discussion on the theoretical proposals for THz QCL based on AlGaN/GaN, it seems that the closest structure to the best AlGaAs/GaAs performance QCL is the one suggested by Sun et al. [Sun et al. 2005]. This structure is presented in Fig. 13.14(top) and is composed of a three well active region with GaN and barriers with $Al_{0.15}Ga_{0.85}N/GaN$. Injection proceeds via the extraction or relaxation of carriers coming from the previous period. Emission occurs between levels 3 and 2, and the extraction is operated by emission of LO phonon from level 2 to 1. The difference in energy between levels 3 to 2 is supposed to be 28 meV that is 6.8 THz. This structure has been designed for a growth on a c-plane taking into account of both spontaneous and piezoelectric fields. Three different scattering processes, LO-phonon emission, LA phonon emission and electron-electron scattering are considered, and the authors succeeded to simulate the threshold current density as a function of lattice temperature. Population inversion is feasible up to room temperature.

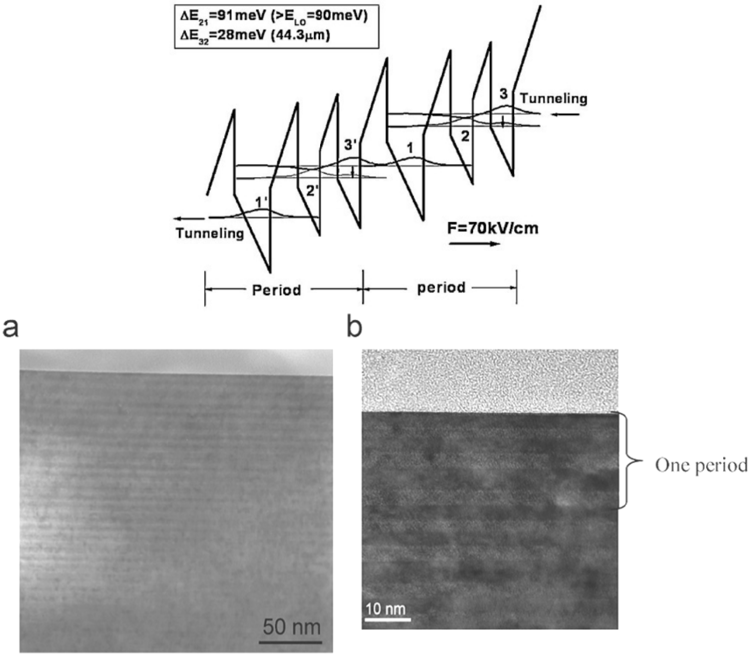


Fig. 13.14. (top) Band structure, subband energy separations and envelope wavefunctions of the active region of the proposed $Al_{0.15}Ga_{0.85}N$/GaN THz QCL. Two periods are shown with each period consisting of 3 GaN QWs and 3 $Al_{0.15}Ga_{0.85}N$ barriers with layer thicknesses (Å): ***30****/40/****30****/25/****20****/25 (barriers in bold and wells in plain) under an electric bias of 70 kV/cm. (bottom) (a) Cross-sectional TEM images of 20 periods MQW sample. (b) Two periods with each period consisting of three GaN QWs and three $Al_{0.20}Ga_{0.80}N$/GaN barriers with layer thickness (Å):* ***30****/40/****30****/25/****20****/25 (barriers in bold and wells in plain). [Adapted with permission from Microelectronics Journal Vol.36, G. Sun, R.A. Soref, "Design and simulation of a GaN/AlGaN quantum cascade laser for terahertz emission," fig. 1, p. 451, Copyright 2005, Elsevier Ltd. and from Journal of Crystal Growth Vol.298 G.S. Huang, T.C. Lu, H.H. Yao, H.C. Kuo, S.C. Wang, G. Sun, C.W. Lin, L. Chang, R.A. Soref," GaN/AlGaN active regions for terahertz quantum cascade lasers grown by low-pressure metal organic vapor deposition," fig. 3, p. 689, Copyright 2007, Elsevier Ltd.]*

A first attempt of an $Al_{0.2}Ga_{0.8}N$/GaN THz QCL has been realized by MOCVD growth [Huang et al. 2007]. The structure grown is close to the one presented by Sun et al, only the aluminum content is modified (from 15% to 20%), the layer thicknesses are the same. The material growth characterization confirms that the sample has a high crystalline quality. In Fig. 13.14(bottom), the TEM pictures show the sharpness of the interfaces of the multi quantum well structure. The authors report an LO phonon energy shift due to the periodic superlattice structure of the active region of 17 cm^{-1} in the QCL structures, compared to that of the $Al_{0.2}Ga_{0.8}N$ film. This value is close to the calculated shift of 19 cm^{-1}. Unfortunately, the authors do not report any electrical measurements, or any emission spectra. This approach seems the most promising one, in order to arrive at a THz room temperature emitter with high output power.

13.9. Summary

Driven by the vast range of applications, THz technologies have attracted the interest of a broad community. However, due to both theoretical and technical difficulties, a compact and efficient THz sources working at room temperature are still lacking. This chapter gives a comprehensive review of the past and current status of the THz technologies.

We presented several examples for the application of THz radiation in terms of THz spectroscopy, T-ray imaging, and using THz as a research tool. Then the existing broadband and narrowband THz sources are discussed. In particular, as a narrowband THz source, the quantum cascade terahertz lasers are drawing much attention. THz QCLs based on both the GaAs and InP material systems are discussed with the performance comparison of different designs. To date, the best performance is achieved with a bound-to-bound LO phonon depopulation design.

Magnetic field has been proven to be an effective tool not only in the investigation scattering mechanisms, but also in improving the performance of a THz QCL. With a strong magnetic field, an enhancement of 45 K of the maximum operating temperature has been demonstrated. In combination with the best bound-to-bound LO phonon depopulation design, a record operating temperature of 225 K was achieved.

The only reported room temperature semiconductor THz emitter is the one based on the difference frequency generation of mid-infrared QCLs. As a second order process, the efficiency is rather low, resulting a THz output power less than 1 μW.

Due to a much bigger optical phonon energy, THz QCLs based on III-nitride material system have been proposed. In this case, the downside of the narrow-gap material system can be partially circumvented, which opens up the possibility of room temperature operation with high output power. Further investigation needs to be carried out to realize a THz emitter in this material system.

References

Ajili, L., Scalari, G., Hoyler, N., Giovannini, M., and Faist, J., “InGaAs–AlInAs/InP terahertz quantum cascade laser,” *Applied Physics Letters* 87, pp. 141107-1, 2005.

Alton, J., Barbieri, S., Fowler, J., Beere, H., Muscat, J., Linfield, E., Ritchie, D., Davies, G., Köhler, R., and Tredicucci, A., “Magnetic field in-plane quantization and tuning of population inversion in a THz superlattice quantum cascade laser,” *Physics Review B* 68, pp.081303-1, 2003.

Altukhov, I.V., Chirkova, E.G., Kagan, M.S., Korolev, K.A., Sinis, V.P., and Yassievich, I.N., “Far-Infrared Stimulated Emission in *p*-Ge under High Uniaxial Pressure,” *Physica Status Solidi* B198 pp. 35, 1996.

Baumann, E., Giorgetta, F., Hofstetter, D., Wu, H., Schaff, W., Eastman, L., and Kirste, L., “Tunneling effects and intersubband absorption in AlN/GaN superlattices,” *Applied Physics Letters* 86, pp. 32110-1, 2005.

Baumann, E., Giorgetta, F.R., Hofstetter, D., Golka, S., Scherenk, W., Strasser, G., Kirste, L., Nicolay, S., Feltin, E. Carlin, J.F., and Grandjean, N., “Near infrared absorption and room temperature photovoltaic response in AlN/GaN superlattices grown by metal-organic vapor-phase epitaxy,” *Applied Physics Letters* 89, pp. 041106-1, 2006.

Becker, C., Vasanelli, A., Sirtori, C., and Bastard, G., “Electron–longitudinal optical phonon interaction between Landau levels in semiconductor heterostructures,” *Physics Review* B 69, pp. 115328-1 (2004)

Belkin, M.A., Capasso, F., Xie, F., Belyanin, A., Fischer, M., Wittmann, A., and Faist, J., “Room temperature terahertz quantum cascade laser source based on intracavity difference-frequency generation,” *Applied Physics Letters* 92, pp. 201101-1, 2008.

Blaser, S., Rochat, M., Beck, M., Hofstetter, D., and Faist, J., “Terahertz intersubband emission in strong magnetic fields,” *Applied Physics Letters* 81, pp. 67, 2002.

Brown, E., McIntosh, K., Nichols, K., and Dennis, C., “Photomixing up to 3.8 THz in low-temperature-grown GaAs,” *Applied Physics Letters* 66, pp. 285, 1995.

Brucherseifer, M., Nagel, M., Haring, B.P., Kurz, H., Bosserhoff A., and Büttner, R., “Label-free probing of the binding state of DNA by time-domain terahertz sensing,” *Applied Physics Letters* 77, pp. 4049, 2000.

Bründermann, E., Chanberlin, D.R., and Haller, E.E., “High duty cycle and continuous terahertz emission from germanium,” *Applied Physics Letters* 76, pp. 2991, 2000.

Bykhovski, A.D., Gelmont, B.L. and M.S. Shur. “Elastic strain relaxation and piezoeffect in GaN-AlN, GaN-AlGaN and GaN-InGaN superlattices,” *Journal of Applied Phycis* 81, pp. 6332, 1997.

Clery, D., “Brainstorming Their Way to an Imaging Revolution,” *Science* 297, pp. 5582, 2002.

Colombelli, R., Capasso, F., Gmachl, C., Hutchinson, A.L., Sivco, D.L., Tredicucci, A., Wanke, M.C., Sergent, A.M. and Cho, A.Y., “Far-infrared surface-plasmon quantum-cascade lasers at 21.5 µm and 24 µm wavelengths,” *Applied Physics Letters* 78, pp. 2620, 2001.

Dyakonov, M. and Shur, M.S. “Shallow water analogy for a ballistic field effect transistor: New mechanism of plasma wave generation by dc current,” *Physics Review Letters* 71, pp. 2465, 1993.

Dyakonov, N., El Fatimy, A., Lusakowski, J., Knap, W., Dyakonov, M.I., Poisson M.A., Morvan E., Bollaert, S., Shchepetov, A., Roelens, Y., Gaquiere, Ch., Theron, D., and Cappy, A., “Room-temperature terahertz emission from nanometer field-effect transistors,” *Applied Physics Letters* 88, pp. 141906-1, 2006.

Fattinger, Ch., and Grischkowsky, D., “Point source terahertz optics,” *Applied Physics Letters* 53, pp. 1480, 1988.

Ferreira, R., “Resonances in the hopping probability between flexible quantum dots: The case of superlattices under parallel electric and magnetic fields,” *Physics Review B* 43, pp. 9336, 1991.

Friel, I., Driscoll, K., Kulenica, E., Dutta, M., Paiella, R., and Moustakas, T., “Investigation of the design parameters of AlN/GaN multiple quantum wells grown by molecular beam epitaxy for intersubband absorption,” *Journal of Crystal Growth* 278, pp. 387, 2005.

Gmachl, C., Ng, H.S., Chu, N.G., and Cho, A.Y. “Intersubband absorption at λ ~1.55 µm in well- and modulation-doped GaN/AlGaN multiple quantum wells with superlattice barriers,” *Applied Physics Letters* 77, 3722 (2000).

Guo, R., Akiyama, K., Minamide, H., and Ito, H., “All-solid-state, narrow linewidth, wavelength-agile terahertz-wave generator,” *Applied Physics Letters* 88, pp. 091120-1, 2006.

Hamazaki, J., Ans, S.M., Kunugita, H., Ema, K., Kanazawa, H., Tachibana, T., Kikuchi, A., and Kishino, K., “Ultrafast intersubband relaxation and nonlinear susceptibility at 1.55 µm in GaN/AlN multiple-quantum wells,” *Applied Physics Letters* 84, pp. 1102, 2004.

Hofstetter, D., Schad, S.S., Wu, H., Shaff, W., and Eastman, L., “GaN/AlN-based quantum-well infrared photodetector for 1.55 µm,” *Applied Physics Letters* 83, pp. 572, 2003.

Huang, G.S., Lu, T.C., Yao, H.H., Kuo, H.C., Wang, S.C., Sun, G., Lin, C., Chang, L., Soref, R.A., “GaN/AlGaN active regions for terahertz quantum cascade lasers grown by low-pressure metal organic vapor deposition,” *Journal of Crystal Growth* 298, pp. 687, 2007.

Humphreys, K., Loughran, J.P., Gradziel, M., Lanigan, W., Ward, T., Murphy, J.A., O’Sullivan, C., Medical applications of terahertz imaging: a review of current technology and potential applications in biomedical engineering,” Proceedings of the 26th Annual International Conference of the IEEE EMBS, San Francisco, CA, USA, pp. 1302-1305, 2004.

Iizuka, N., Kaneko, K., Suzuki, N., Asano, T., Noda, S., and Wada, O., “Ultrafast intersubband relaxation (<=150 fs) in AlGaN/GaN multiple quantum wells,” *Applied Physics Letters* 77, pp. 648, 2000.

Inguscio, M., Moruzzi, G., Evenson, K.M., and Jennings, D.A., "A review of frequency measurements of optically pumped lasers from 0.1 to 8 THz," *Journal of Applied Physics* 60, pp. R161, 1986.

Jovanović, V.D., Indjin, D., Ikonić, Z., and Harrison, P., "Simulation and design of GaN/AlGaN far-infrared (λ ~34 μm) quantum-cascade laser," *Applied Physics Letters* 84, pp. 2995, 2004.

Jukam, N., Dhillon, S.S., Oustinov, D., Zhao, Z.Y., Hameau, S., Tignon, J., Barbieri, S., Vasanelli, A., Filloux, P., Sirtori, C., and Marcadet, X., "Investigation of spectral gain narrowing in quantum cascade lasers using terahertz time domain spectroscopy," *Applied Physics Letters* 93, pp. 101115-1, 2008.

Julien, F.H., Tchernycheva, M., Nevou, L., Doyennette, L., Colombelli, R., Warde, E., Guillot, F., and Monroy, E., "Nitride intersubband devices: prospects and recent developments," *Physica Status Solidi. (a)* 204, pp. 1987, 2007.

Kagan, M.S., Altukhov, I.V., Sinis, V.P., Thomas, S.G., Wang, K.L., Chao, K.A., and Yassievich, I.N., "Terahertz emission of SiGe/Si quantum wells," *Thin Solid Films* 380, pp. 237, 2000.

Knap, W., Lusakowski, J., Parenty, T., Bollaert, S., Cappy, A., Popov, V.V., and Shur, M.S., "Terahertz emission by plasma waves in 60 nm gate high electron mobility transistors," *Applied Physics Letters* 84, pp. 2331, 2004.

Kempa, K., Zhou, Y., Engelbrecht, J., Bakshi, P., Ha, H., Moser, J., Naughton, M., Ulrich, J., Strasser, G., Gornik, E., and Unterrainer, K., "Intersubband Transport in Quantum Wells in Strong Magnetic Fields Mediated by Single- and Two-Electron Scattering," *Physics Review Letters* 88, pp. 226803, 2002.

Kishino, K., Kikuchi, A., Kanazawa, H., and Tachibana, T., "Intersubband transition in $(GaN)_m/(AlN)_n$ superlattices in the wavelength range from 1.08 to 1.61 μm," *Applied Physics Letters* 81, pp. 1234, 2000.

Köhler, R., Tredicucci, A., Beltram, F., Beere, H.E., Linfield, E.H., Davies, A.G., Ritchie, D.A., Dhillon, S.S., and Sirtori, C., "High-performance continuous-wave operation of superlattice terahertz quantum-cascade lasers," *Applied Physics Letters* 82, pp. 1518, 2003.

Kröll, J., Darmo, J., Dhillon, S.S., Marcadet, X., Calligaro, M., Sirtori, C., and Unterrainer, K., "Phase-resolved measurements of stimulated emission in a laser," *Nature* 449, pp 698, 2007.

Kumar, S., Hu, Q., and Reno, J.L., "186 K operation of terahertz-quantum cascade lasers based on diagonal design," *Applied Physics Letters*, 94 pp.131105, 2009.

Leuliet, A., Vasanelli, A., Wade, A., Fedorov, G., Smirnov, D., Bastard, G., and Sirtori, C., "Electron scattering spectroscopy by a high magnetic field in quantum cascade lasers," *Physics Review B* 73, pp. 085311-1, 2006.

Mittleman, D.M., Jacobsen, R.H., and Nuss, M.C., "T-ray imaging," *IEEE Journal of Selected Topics in Quantum Electronics* 2, pp. 679, 1996.

Nagel, M., Bolivar, P.H., Brucherseifer, M., Kurz, H., Bosserhoff, A., and Büttner, R., "Integrated THz technology for label-free genetic diagnostics," *Applied Physics Letters* 80, pp. 154, 2002.

Nevou, L., Tchernycheva, M., Julien, F.H., Guillot, F., and Monroy, E., "Short wavelength (λ =2.13 μm) intersubband luminescence from GaN/AlN quantum wells at room temperature," *Applied Physics Letters* 90, pp. 121106-1, 2007.

Nicolay, S., Feltin, E., Carlin, J.F., Grandjean, N., Nevou, L., Julien, F.H., Schmidbauer, M., Remmele, T., and Albrecht, M., "Strain-induced interface instability in GaN/AlN multiple quantum wells," *Applied Physics Letters* 91, pp. 061927-1, 2007.

Nishizawa, J., Tanabe, T., Suto, K., Watanabe, Y., Sasaki T., and Oyama, Y., "Continuous-Wave Frequency-Tunable Terahertz-Wave Generation From GaP," *IEEE Photonics Technology Letters* 18, pp. 2008, 2006.

Pavlov, S.G., Zhukavin, R.K., Orlova, E.E., Shastin, V.N., Kirsanov, A.V., Hubers, H.W., and Auen, K., "Stimulated Emission from Donor Transitions in Silicon," *Physics Review Letters* 84, pp. 5220, 2000.

Péré-Laperne, N., Vaulchier, L.A., Guldner, Y., Bastard, G., Scalari, G., Giovannini, M., Faist, J., Vasanelli, A., Dhillon, S., and Sirtori, C., "Inter-Landau level scattering and LO-phonon emission in terahertz quantum cascade laser," *Applied Physics Letters* 91, pp. 062102-1, 2007.

Pflüg, C., Belkin, M.A., Wang, Q.J., Geiser, M., Belyanin, A., Fischer, M., Wittmann, A., Faist, J., and Capasso, F., "Surface-emitting terahertz quantum cascade laser source based on intracavity difference-frequency generation," *Applied Physics Letters* 93, pp. 161110-1, 2008.

Phillips, T.G. andKeene, J., "Submillimeter astronomy," *Proceedings of the IEEE* 80, pp. 1662, 1992.

Prasad, T., Colvin, V.L., and Mittleman, D.M., "The effect of structural disorder on guided resonances in photonic crystal slabs studied with terahertz time-domain spectroscopy," *Optics Express* 15, pp. 16954-1, 2007.

Rochat, M., Ajili, L., Willenberg, H., Faist, J., Beere, H., Davies G., Linfield, E., and Ritchie, D., "Low-threshold terahertz quantum-cascade lasers," *Applied Physics Letters* 81, pp. 1381, 2002.

Scalari, G., Ajili, L., Fais, J., Beere, H., Linfield, E., Ritchie, D., and Davies, G., "Far-infrared ($\lambda \sim 87$ μm) bound-to-continuum quantum-cascade lasers operating up to 90 K," *Applied Physics Letters* 82, pp. 3165, 2003a.

Scalari, G., Blaser, S., Ajili, L., Faist, J., Beere, H., Linfield, E., Ritchie, D., and Davies, G., "Population inversion by resonant magnetic confinement in terahertz quantum-cascade lasers," *Applied Physics Letters* 83, pp. 3453, 2003b.

Scalari, G., Blaser, S., Faist, J., Beere, H., Linfield, E., Ritchie, D., and Davies, G., "Terahertz Emission from Quantum Cascade Lasers in the Quantum Hall Regime: Evidence for Many Body Resonances and Localization Effects," *Physics Review Letters* 93,pp. 237403, 2004.

Scalari, G., Walther, C., Faist, J., Beere, H., and Ritchie, D., "Electrically switchable, two-color quantum cascade laser emitting at 1.39 and 2.3 THz," *Applied Physics Letters* 88, pp. 141102-1, 2006.

Scalari, G., Walther, C., Fischer, M., Terazzi, R., Beere, H., Ritchie, D., and Faist, J., "THz and sub-THz quantum cascade lasers," *Laser & Photonics Review* 3, pp. 45, 2008.

Shi, W.,, Ding, Y.J., Fernelius, N., and Vodopyanov, K., "Efficient, tunable, and coherent 0.18–5.27-THz source based on GaSe crystal," *Optics Letters* 27, pp. 1454, 2002

Siegel, P.H., "Terahertz technology," *IEEE Transactions on Microwave Theory and Techniques* 50, pp. 910, 2002.

Smirnov, D., Becker, C., Drachenko, O., Rylkov, V.V., Page, H., Leotin,v, and Sirtori, C., "Control of electron–optical-phonon scattering rates in quantum box cascade lasers," *Physics Review B* 66, pp. 121305-1, 2002.

Smirnov, D., Drachenko, O., Leotin, J., Page H., Becker, C., Sirtori C., Apalkov, V., and Chakraborty, T., "Intersubband magnetophonon resonances in quantum cascade structures," *Physics Review B* 66, pp. 125317, 2002.

Sun, G., Soref, R.A., Microelectronics Journal 36, pp. 450, 2005.

Suto, K., Sasakia, T., Tanabe, T., Saito, K., Nishizawa, J., and Ito, M., "GaP THz wave generator and THz spectrometer using Cr:Forsterite lasers," *Review of Scientific Instruments* 76, pp. 123109, 2005.

Suzuki, N., Iizuka, N., and Kaneko, K., "Calculation of Near-Infrared Intersubband Absorption Spectra in GaN/AlN Quantum Wells," *Japaneese Journal of Applied Physics* 42, pp. 132, 2002.

Tanabe, T., Suto, K., Nishizawa, J., Saito, K., and Kimura, T., "Tunable terahertz wave generation in the 3- to 7-THz region from GaP," *Applied Physics letters* 83, pp. 237, 2003.

Unterrainer, K., Kremser, C., Gornik, E., Pidgeon, C.R., Ivanov, Y.L., and Haller, E.E., "Tunable cyclotron-resonance laser in germanium," *Physics Review Letters* 64, pp. 2277, 1990.

Vasanelli, A., Leuliet, A., Sirtori, C., Wade, A., Fedorov, G., Smirnov, D., and Bastard, G., "Role of elastic scattering mechanisms in GaInAs/AlInAs quantum cascade lasers," *Applied Physics Letters* 89, pp. 172120-1, 2006.

Vitiello, M.S., Scamarcio, G., Spagnolo, V., Williams, B., Kumar, S., Hu, Q., and Reno, J., "Measurement of subband electronic temperatures and population inversion in THz quantum-cascade lasers," *Applied Physics Letters* 86, pp. 111115-1, 2005.

Vukmirović, N., Jovanović, V.D., Indjin, D., Ikonić, Z., Harrison, P., and Milanović, V., "Optically pumped terahertz laser based on intersubband transitions in a GaN/AlGaN double quantum well," *Journal of Applied Phycis* 97, pp. 103106-1, 2005.

Wade,A., Fedorov, G., Smirnov, D., Kumar, S., Williams, B.S., Hu, Q., and Reno, J.L., "Magnetic-field-assisted terahertz quantum cascade laser operating up to 225 K," *Nature Photonics* 3, pp. 41, 2008.

Waki, I., Kumtornkittikul, C., Shimogaki, Y., and Nakano, Y., "Shortest intersubband transition wavelength (1.68 μm) achieved in AlN/GaN multiple quantum wells by metalorganic vapor phase epitaxy," *Applied Physics Letters* 82, pp. 4465, 2003

Waki, I., Kumtornkittikul, C., Shimogaki, Y., and Nakano, Y., "Erratum: "Shortest intersubband transition wavelength (1.68 μm) achieved in AlN/GaN multiple quantum wells by metalorganic vapor phase epitaxy" [Appl. Phys. Lett. 82, 4465 (2003)]," *Applied Physics Letters* 84, pp. 3703, 2003.

Williams, B.S., Callebaut, H., Kumar, S., Hu, Q., and Reno, J.L., "3.4-THz quantum cascade laser based on longitudinal-optical-phonon scattering for depopulation," *Applied Physics Letters* 82, pp. 1015, 2003.

Williams, B.S., Kumar, S., Callebaut, H., Hu, Q., and Reno, J.L., "Terahertz quantum-cascade laser operating up to 137 K," *Applied Physics Letters* 83, pp. 5142, 2003.

Williams, B.S., Kumar, S., Callebaut, H., Hu, Q., and Reno, J.L., "Terahertz quantum-cascade laser at $\lambda \sim 100$ μm using metal waveguide for mode confinement," *Applied Physics Letters* 83, pp. 2124, 2003.

Williams, B.S., "Terahertz quantum-cascade lasers," *Nature photonics* 1, pp. 517, 2007.

Woolard, D.L., Globus, T.R., Gelmont, B.L., Bykhovskaia, M., Samuels, A.C., Cookmeyer, D., Hesler, J.L., Crowe, T.W., Jensen, J.O., Jensen, J.L., and Loerop, W.R., "Submillimeter-wave phonon modes in DNA macromolecules," *Physics Review E* 65, pp. 051903, 2002.

Woolard, D.L., Brown, E.R., Samuels, A.C., Jensen, J.O., Globus, T., Gelmont, B., and Wolski, M., "Terahertz-frequency remote-sensing of biological warfare agents," *IEEE MTT-S Digest,* pp. 763, 2003.

Further reading

Miles, R.E., Harrison, P., and Lippens, D., *Terahertz Sources and Systems*, Springer, 2001

Kiyomi Sakai and Jōhō Tsūshin Kenkyū Kikō, *Terahertz optoelectronics*, Birkhäuser, 2005.

Redo-Sanchez, A., and Xi-Cheng Zhang, "Terahertz Science and Technology Trends," *IEEE Journal of Selected Topics in Quantum Electronics* 14, pp. 260, 2008.

Kumar, S., and Lee, A.W.M., "Resonant-Phonon Terahertz Quantum-Cascade Lasers and Video-Rate Terahertz Imaging," *IEEE Journal of Selected Topics in Quantum Electronics* 14, pp. 333, 2008.

Problems

1. Suppose that the wavelength of a laser radiation is denoted by λ, and the unit is μm. The photon energy is denoted by ε with a unit of meV. The frequency is denoted by f with a unit of THz. A simple relation among the nominal numerical value of them would be very helpful when one needs to convert one into another. An example is that the product of λ and ε is approximately 1240. Find a similar relation between λ and f. Use this relation to tell the wavelength of a THz laser whose frequency is 3.5 THz and the frequency of a THz laser whose wavelength is 120 μm.

2. We know that for InP based mid-infrared QCLs, the waveguide is composed of two bulk cladding layers located below and above the laser core. A confined optical mode exists when the refractive index of the cladding is smaller than that of the laser core. A rule of thumb is that the total thickness of the epitaxial structure is comparable to the wavelength. For example, the epitaxial thickness of a 5 μm QCL is about 6 μm and that of a 10 μm QCL is about 10 μm. Explain why it is not possible to confine the optical mode of a THz QCL in a way similar to the mid-infrared case. Discussion should be addressed for both GaAs

and InP based material systems in terms of the choice of the cladding material and the epitaxial thickness.

3. The emission wavelength of a laser λ, the operating temperature T, and strength of the magnetic field B, these drastically different physical quantities can be compared by converting them into energy. Comparing with to the case in the mid-infrared region, the difficulty of making a working THz laser at room temperature and the convenience of studying a THz laser under the magnetic field can be understood from this perspective. As an example, explain this situation for a mid-infrared laser with a wavelength of 5 μm and a THz laser with a frequency of 3 THz.

3. A GaAs based THz QCL emits at 3.5 THz. The number of QCL stages inside the laser corer is 200. The core design is the bond-to-bond LO phonon depopulation. The LO phonon energy of GaAs is approximately 36 meV. (a). Find the minimum operating voltage for the device. (b). What is the wall plug efficiency of this device if the current is 1 A and the output power is 100 mW?

4. In the mid-infrared case, the dipole matrix element z_{ij} is usually used for the gain calculation, where z_{ij} has a length unit. For the THz case, however, the unit-less oscillator strength f_{ij}, is usually used for the same purpose. In fact, there is a definite relation between the dipole matrix element and the oscillator strength, which requires the knowledge of the transition energy ε_{ij} and the effective mass m^*, i. e, $f_{ij} = \frac{2m^*\varepsilon_{ij}}{\hbar^2}\left|z_{ij}\right|^2$. Calculate the lasing oscillator strength of a 3 THz GaAs based QCL with a dipole matrix element of 5 nm.

5. For mid-infrared QCLs, the most important scattering mechanism is the polar optical phonon (LO phonon) scattering. A typical LO phonon scattering rate for the GaAs based THz QCL is on the order of 1 ps^{-1}, i. e, a scattering lifetime of 1 ps. The polar optical phonon scattering rate $R_{pop} \propto m^*\left(\frac{1}{\varepsilon_\infty} - \frac{1}{\varepsilon_s}\right)$, where m^* is the effective mass, and ε_∞ and ε_s are the high frequency and static dielectric constant, respectively. Show that the LO phonon scattering rate for the GaN based material system is

more than an order of magnitude faster than the GaAs based material system, i.e., a scattering lifetime of less than 100 fs. Use the parameters in the following table for the calculation.

	m^*	ε_∞	ε_s
GaAs	$0.067m_0$	10.9	12.9
GaN	$0.19m_0$	5.8	10.4

Appendices

A.1. Physical constants
A.2. International system of units (SI units)
A.3. Physical properties of elements in the periodic table
A.4. Physical properties of important semiconductors
A.5. Thermionic emission
A.6. Minority carrier lifetime measurement
A.7. Advanced topics in Type-II photodetectors
A.8. Physical properties and safety information of metalorganics

M. Razeghi, *Technology of Quantum Devices*, DOI 10.1007/978-1-4419-1056-1,
© Springer Science+Business Media, LLC 2010

A.1. Physical constants

Angström unit	Å	10^{-10} m = 10^{-8} cm = 10^{-4} µm
Avogadro constant	$\mathcal{N}_A$	6.02204×10^{23} mol^{-1}
Bohr radius	a_0	0.52917 Å
Boltzmann constant	k_b	1.38066×10^{-23} J·K^{-1} (=$R/\mathcal{N}_A$)
		8.61738×10^{-5} eV·K^{-1}
Calorie	cal	4.184 J
Elementary charge	q	1.60218×10^{-19} C
Electron rest mass	m_0	0.91095×10^{-30} kg
Electron Volt	eV	1.60218×10^{-19} J
		23.053 kcal·mol^{-1}
Gravitational constant	g	9.81 m·s^{-2}
Gas constant	R	1.98719 cal·mol^{-1}·K^{-1}
		8.31440 J·mol^{-1}·K^{-1}
Permeability in vacuum	μ_0	$4\pi10^{-9} = 1.25633\times10^{-6}$ H·m^{-1}
Permittivity in vacuum	ε_0	8.85418×10^{-12} F·m^{-1}
		(=$1/\mu_0c^2$)
Plank's constant	h	6.62617×10^{-34} J·s
Reduced Plank's constant	$\hbar$	1.05458×10^{-34} J·s (=$h/2\pi$)
Proton rest mass	M_p	1.67264×10^{-27} kg
Standard atmosphere	atm	1.01325×10^{5} N·m^{-2}
Thermal voltage at 300 K	k_bT/q	0.0259 V
Velocity of light in vacuum	c	2.99792×10^{8} m·s^{-1}
Wavelength of 1 eV quantum	λ	1.23977 µm

A.2. International system of units (SI units)

Base units

Quantity	*Unit name*	*Unit symbol*
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

Prefixes

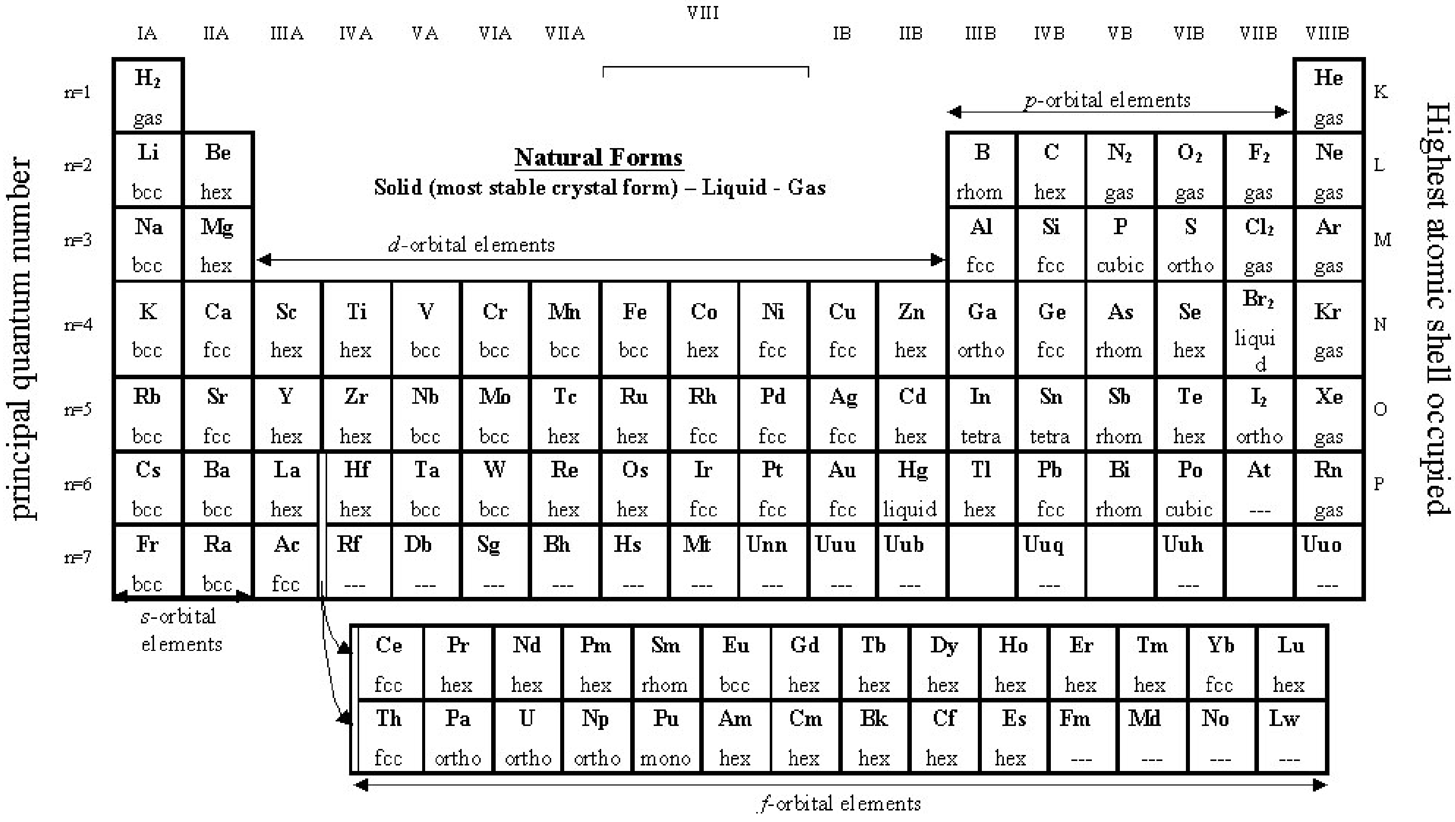
Factor	*Prefix*	*Symbol*	*Factor*	*Prefix*	*Symbol*
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	µ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a
10^{2}	hecto	h	10^{-21}	zepto	z
10^{1}	deka	da	10^{-24}	yocto	y

Derived units

Quantity	*Special name*	*Unit Symbol*	*Dimension*
Angle	radian	-	rad
Solid angle	steradian	-	sr
Speed, velocity	-	-	$m\cdot s^{-1}$
Acceleration	-	-	$m\cdot s^{-2}$
Angular velocity,	frequency	-	$rad\cdot s^{-1}$
Angular acceleration	-	-	$rad\cdot s^{-2}$
Frequency	hertz	Hz	s^{-1}
Force	newton	N	$kg\cdot m\cdot s^{-2}$
Pressure, stress	pascal	Pa	$N\cdot m^{-2}$
Work, energy, heat	joule	J	$N\cdot m$, $kg\cdot m^2\cdot s^{-2}$
Power	watt	W	$J\cdot s$
Electric charge	coulomb	C	$A\cdot s$
Electric potential	volt	V	$J\cdot C^{-1}$, $W\cdot A^{-1}$
Resistance	ohm	Ω	$V\cdot A^{-1}$
Conductance	siemens	S	$A\cdot V^{-1}$, Ω^{-1}
Magnetic flux	weber	Wb	$V\cdot s$
Inductance	henry	H	$Wb\cdot A^{-1}$
Capacitance	farad	F	$C\cdot V^{-1}$
Electric field strength	-	-	$V\cdot m^{-1}$, $N\cdot C^{-1}$
Magnetic induction	tesla	T	$Wb\cdot m^{-2}$, $N.A^{-1}\cdot m^{-1}$
Electric displacement	-	-	$C\cdot m^{-2}$
Magnetic field strength	-	-	$A\cdot m^{-1}$
Celsius temperature	degrees Celsius	°C	K
Luminous flux	lumen	lm	$cd\cdot sr$
Illuminance	lux	lx	$lm\cdot m^{-2}$
Radioactivity	becquerel	Bq	s^{-1}
Catalytic activity	katal	kat	$mol\cdot s^{-1}$

A.3. Physical properties of elements in the periodic table

The following figures summarize the general physical properties of most elements in the periodic table. These include their natural forms (Fig. A.1) with the structure in which they crystallize, their density of mass (Fig. A.2), boiling point (Fig. A.3), melting point (Fig. A.4), thermal conductivity (Fig. A.5), molar volume (Fig. A.6), specific heat (Fig. A.7), atomic radius (Fig. A.8), oxidation states (Fig. A.9), ionic radius (Fig. A.10), electronegativity (Fig. A.11), and electron affinity (Fig. A.12).



Natural Forms

Solid (most stable crystal form) – Liquid - Gas

p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	Highest atomic shell occupied
n=1	**H_2** gas																	**He** gas	K
n=2	**Li** bcc	**Be** hex											**B** rhom	**C** hex	**N_2** gas	**O_2** gas	**F_2** gas	**Ne** gas	L
n=3	**Na** bcc	**Mg** hex											**Al** fcc	**Si** fcc	**P** cubic	**S** ortho	**Cl_2** gas	**Ar** gas	M
n=4	**K** bcc	**Ca** fcc	**Sc** hex	**Ti** hex	**V** bcc	**Cr** bcc	**Mn** bcc	**Fe** bcc	**Co** hex	**Ni** fcc	**Cu** fcc	**Zn** hex	**Ga** ortho	**Ge** fcc	**As** rhom	**Se** hex	**Br_2** liquid	**Kr** gas	N
n=5	**Rb** bcc	**Sr** fcc	**Y** hex	**Zr** hex	**Nb** bcc	**Mo** bcc	**Tc** hex	**Ru** hex	**Rh** fcc	**Pd** fcc	**Ag** fcc	**Cd** hex	**In** tetra	**Sn** tetra	**Sb** rhom	**Te** hex	**I_2** ortho	**Xe** gas	O
n=6	**Cs** bcc	**Ba** bcc	**La** hex	**Hf** hex	**Ta** bcc	**W** bcc	**Re** hex	**Os** hex	**Ir** fcc	**Pt** fcc	**Au** fcc	**Hg** liquid	**Tl** hex	**Pb** fcc	**Bi** rhom	**Po** cubic	**At** ---	**Rn** gas	P
n=7	**Fr** bcc	**Ra** bcc	**Ac** fcc	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce fcc	**Pr** hex	**Nd** hex	**Pm** hex	**Sm** rhom	**Eu** bcc	**Gd** hex	**Tb** hex	**Dy** hex	**Ho** hex	**Er** hex	**Tm** hex	**Yb** fcc	**Lu** hex
Th fcc	**Pa** ortho	**U** ortho	**Np** ortho	**Pu** mono	**Am** hex	**Cm** hex	**Bk** hex	**Cf** hex	**Es** hex	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.1. Natural forms of elements in the periodic table.

Density: solid, liquid in g.cm^{-3} (20°C, 1 atm)
gas in g.l^{-1} (0°C, 1 atm)

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	Highest atomic shell occupied
n=1	H_2 0.089																	He 0.179	K
n=2	Li 0.53	Be 1.85											B 2.34	C 2.62	N_2 1.25	O_2 1.43	F_2 1.70	Ne 0.90	L
n=3	Na 0.97	Mg 1.74											Al 2.70	Si 2.33	P 1.82	S 2.07	Cl_2 3.21	Ar 1.78	M
n=4	K 0.86	Ca 1.55	Sc 3.0	Ti 4.54	V 6.11	Cr 7.19	Mn 7.44	Fe 7.87	Co 8.90	Ni 8.90	Cu 8.96	Zn 7.13	Ga 5.90	Ge 5.32	As 5.72	Se 4.80	Br_2 3.12	Kr 3.73	N
n=5	Rb 1.63	Sr 2.54	Y 4.50	Zr 6.50	Nb 8.57	Mo 10.2	Tc 11.5	Ru 12.4	Rh 12.4	Pd 12.0	Ag 10.5	Cd 8.65	In 7.31	Sn 7.3	Sb 6.69	Te 6.24	I_2 4.92	Xe 5.89	O
n=6	Cs 1.87	Ba 3.5	La 6.15	Hf 13.31	Ta 16.6	W 19.3	Re 21.0	Os 22.6	Ir 22.4	Pt 21.5	Au 19.3	Hg 13.55	Tl 11.85	Pb 11.4	Bi 9.8	Po 9.3	At ---	Rn 9.73	P
n=7	Fr ---	Ra 5.0	Ac 10.07	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

s-orbital elements

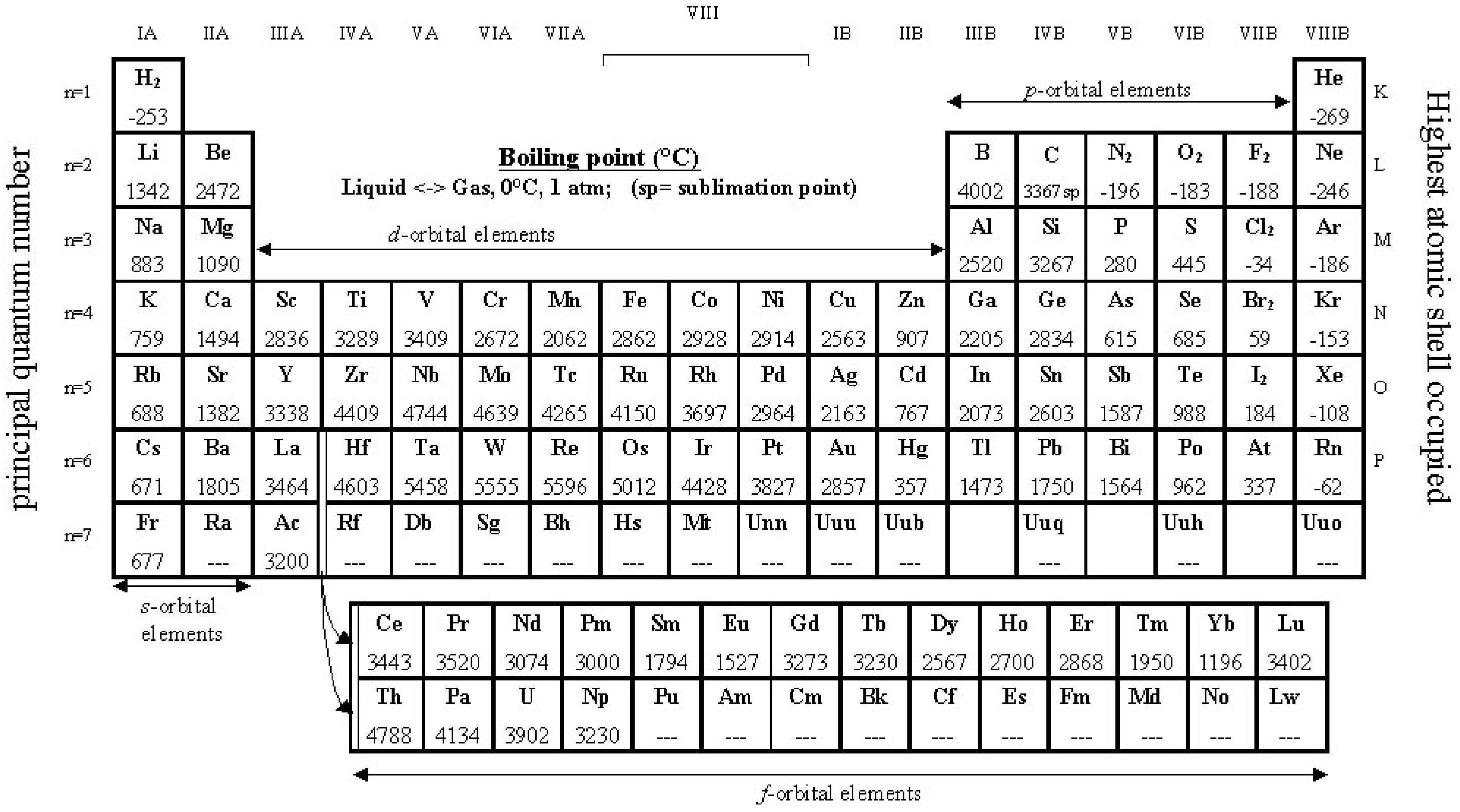
d-orbital elements

p-orbital elements

Ce 6.77	Pr 6.77	Nd 7.01	Pm 7.30	Sm 7.52	Eu 5.24	Gd 7.90	Tb 8.23	Dy 8.55	Ho 8.80	Er 9.07	Tm 9.32	Yb 6.90	Lu 9.84
Th 11.7	Pa 15.4	U 18.95	Np 20.3	Pu 19.8	Am 13.6	Cm 13.5	Bk ---	Cf ---	Es ---	Fm ---	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.2. Density of mass of elements in the periodic table.

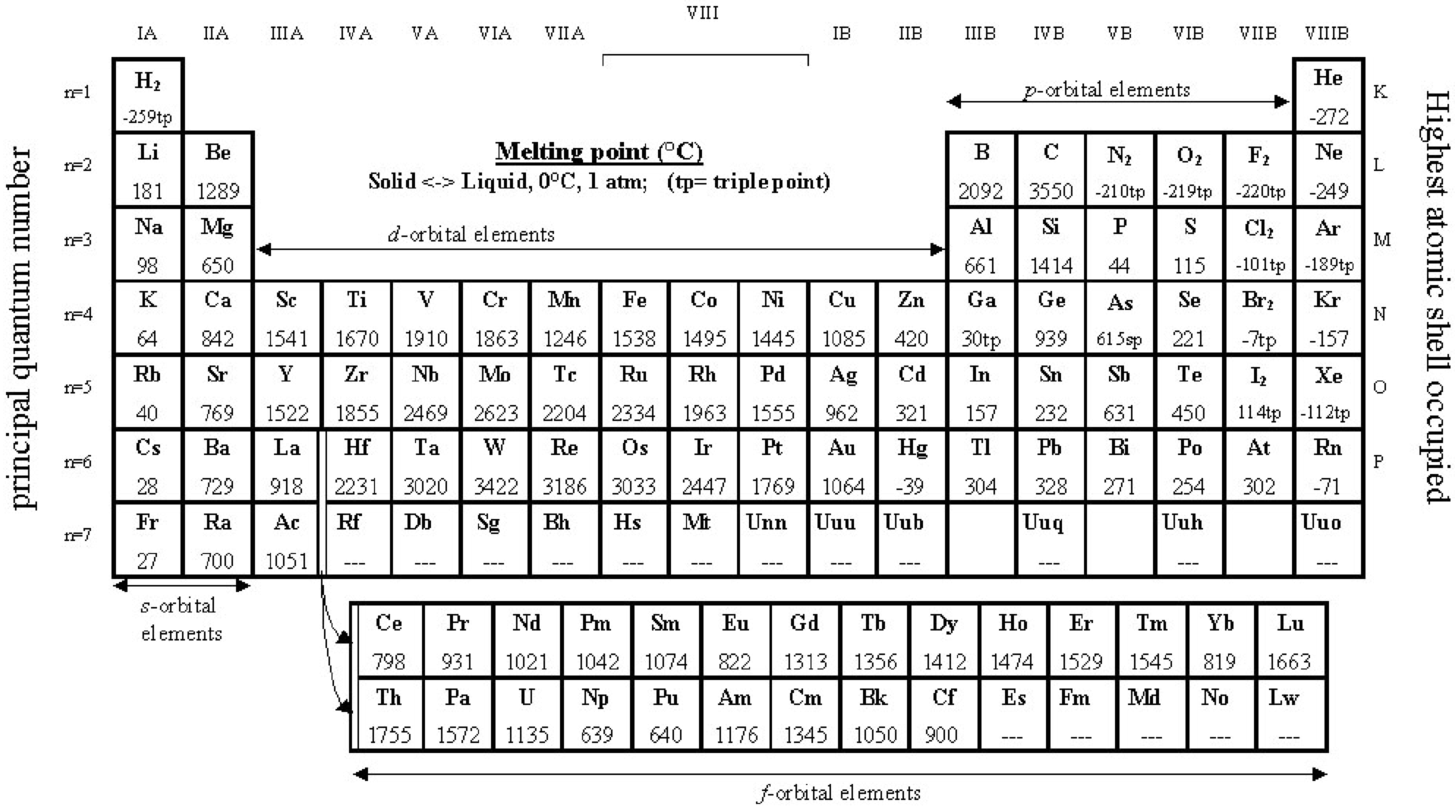

Highest atomic shell occupied
principal quantum number
p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

Boiling point (°C)
Liquid <-> Gas, 0°C, 1 atm; (sp= sublimation point)

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H$_2$** -253																	**He** -269	K
n=2	**Li** 1342	**Be** 2472											**B** 4002	**C** 3367 sp	**N$_2$** -196	**O$_2$** -183	**F$_2$** -188	**Ne** -246	L
n=3	**Na** 883	**Mg** 1090											**Al** 2520	**Si** 3267	**P** 280	**S** 445	**Cl$_2$** -34	**Ar** -186	M
n=4	**K** 759	**Ca** 1494	**Sc** 2836	**Ti** 3289	**V** 3409	**Cr** 2672	**Mn** 2062	**Fe** 2862	**Co** 2928	**Ni** 2914	**Cu** 2563	**Zn** 907	**Ga** 2205	**Ge** 2834	**As** 615	**Se** 685	**Br$_2$** 59	**Kr** -153	N
n=5	**Rb** 688	**Sr** 1382	**Y** 3338	**Zr** 4409	**Nb** 4744	**Mo** 4639	**Tc** 4265	**Ru** 4150	**Rh** 3697	**Pd** 2964	**Ag** 2163	**Cd** 767	**In** 2073	**Sn** 2603	**Sb** 1587	**Te** 988	**I$_2$** 184	**Xe** -108	O
n=6	**Cs** 671	**Ba** 1805	**La** 3464	**Hf** 4603	**Ta** 5458	**W** 5555	**Re** 5596	**Os** 5012	**Ir** 4428	**Pt** 3827	**Au** 2857	**Hg** 357	**Tl** 1473	**Pb** 1750	**Bi** 1564	**Po** 962	**At** 337	**Rn** -62	P
n=7	**Fr** 677	**Ra** ---	**Ac** 3200	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce 3443	**Pr** 3520	**Nd** 3074	**Pm** 3000	**Sm** 1794	**Eu** 1527	**Gd** 3273	**Tb** 3230	**Dy** 2567	**Ho** 2700	**Er** 2868	**Tm** 1950	**Yb** 1196	**Lu** 3402
Th 4788	**Pa** 4134	**U** 3902	**Np** 3230	**Pu** ---	**Am** ---	**Cm** ---	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.3. Boiling point of elements in the periodic table.



Highest atomic shell occupied

principal quantum number

Melting point (°C)

Solid <-> Liquid, 0°C, 1 atm; (tp= triple point)

p-orbital elements

d-orbital elements

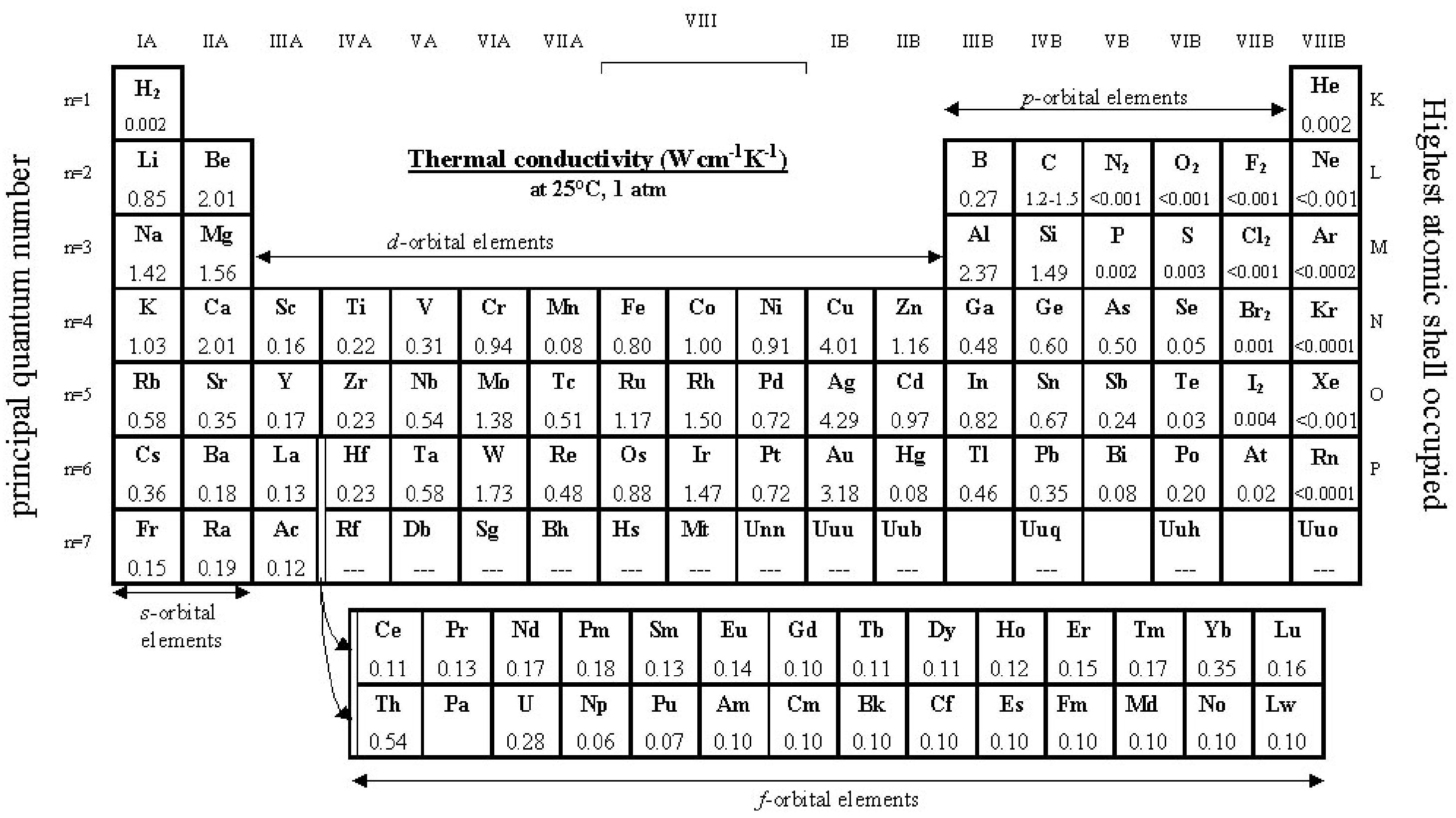
s-orbital elements

f-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H_2** -259tp																	**He** -272	K
n=2	**Li** 181	**Be** 1289											**B** 2092	**C** 3550	**N_2** -210tp	**O_2** -219tp	**F_2** -220tp	**Ne** -249	L
n=3	**Na** 98	**Mg** 650											**Al** 661	**Si** 1414	**P** 44	**S** 115	**Cl_2** -101tp	**Ar** -189tp	M
n=4	**K** 64	**Ca** 842	**Sc** 1541	**Ti** 1670	**V** 1910	**Cr** 1863	**Mn** 1246	**Fe** 1538	**Co** 1495	**Ni** 1445	**Cu** 1085	**Zn** 420	**Ga** 30tp	**Ge** 939	**As** 615sp	**Se** 221	**Br_2** -7tp	**Kr** -157	N
n=5	**Rb** 40	**Sr** 769	**Y** 1522	**Zr** 1855	**Nb** 2469	**Mo** 2623	**Tc** 2204	**Ru** 2334	**Rh** 1963	**Pd** 1555	**Ag** 962	**Cd** 321	**In** 157	**Sn** 232	**Sb** 631	**Te** 450	**I_2** 114tp	**Xe** -112tp	O
n=6	**Cs** 28	**Ba** 729	**La** 918	**Hf** 2231	**Ta** 3020	**W** 3422	**Re** 3186	**Os** 3033	**Ir** 2447	**Pt** 1769	**Au** 1064	**Hg** -39	**Tl** 304	**Pb** 328	**Bi** 271	**Po** 254	**At** 302	**Rn** -71	P
n=7	**Fr** 27	**Ra** 700	**Ac** 1051	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce 798	**Pr** 931	**Nd** 1021	**Pm** 1042	**Sm** 1074	**Eu** 822	**Gd** 1313	**Tb** 1356	**Dy** 1412	**Ho** 1474	**Er** 1529	**Tm** 1545	**Yb** 819	**Lu** 1663
Th 1755	**Pa** 1572	**U** 1135	**Np** 639	**Pu** 640	**Am** 1176	**Cm** 1345	**Bk** 1050	**Cf** 900	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

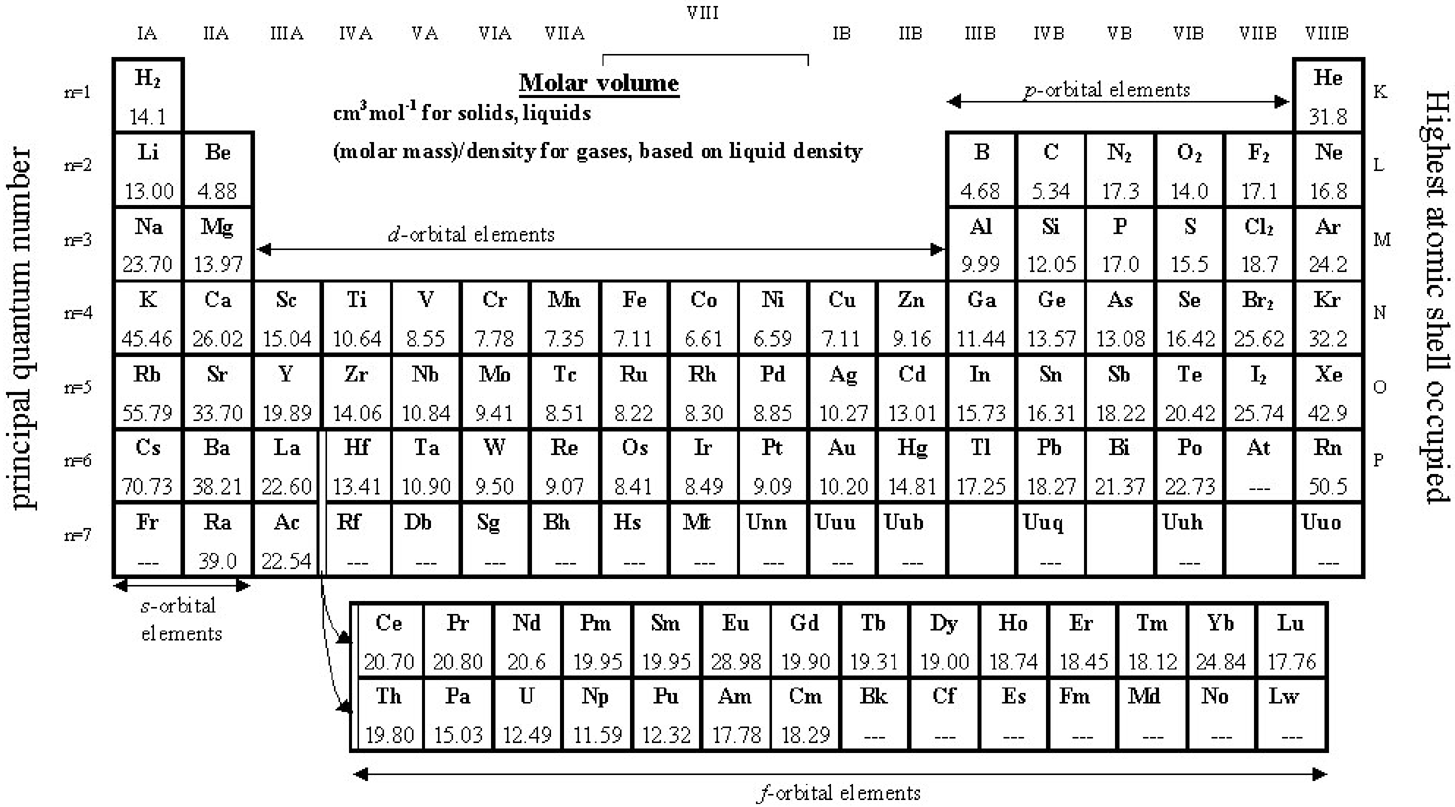
Fig. A.4. Melting point of elements in the periodic table.


Highest atomic shell occupied
principal quantum number
Thermal conductivity (W cm^{-1}K^{-1})
at 25°C, 1 atm
s-orbital elements
d-orbital elements
p-orbital elements
f-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII	VIII	VIII	IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 0.002																	He 0.002	K
n=2	Li 0.85	Be 2.01											B 0.27	C 1.2-1.5	N_2 <0.001	O_2 <0.001	F_2 <0.001	Ne <0.001	L
n=3	Na 1.42	Mg 1.56											Al 2.37	Si 1.49	P 0.002	S 0.003	Cl_2 <0.001	Ar <0.0002	M
n=4	K 1.03	Ca 2.01	Sc 0.16	Ti 0.22	V 0.31	Cr 0.94	Mn 0.08	Fe 0.80	Co 1.00	Ni 0.91	Cu 4.01	Zn 1.16	Ga 0.48	Ge 0.60	As 0.50	Se 0.05	Br_2 0.001	Kr <0.0001	N
n=5	Rb 0.58	Sr 0.35	Y 0.17	Zr 0.23	Nb 0.54	Mo 1.38	Tc 0.51	Ru 1.17	Rh 1.50	Pd 0.72	Ag 4.29	Cd 0.97	In 0.82	Sn 0.67	Sb 0.24	Te 0.03	I_2 0.004	Xe <0.001	O
n=6	Cs 0.36	Ba 0.18	La 0.13	Hf 0.23	Ta 0.58	W 1.73	Re 0.48	Os 0.88	Ir 1.47	Pt 0.72	Au 3.18	Hg 0.08	Tl 0.46	Pb 0.35	Bi 0.08	Po 0.20	At 0.02	Rn <0.0001	P
n=7	Fr 0.15	Ra 0.19	Ac 0.12	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

Ce 0.11	Pr 0.13	Nd 0.17	Pm 0.18	Sm 0.13	Eu 0.14	Gd 0.10	Tb 0.11	Dy 0.11	Ho 0.12	Er 0.15	Tm 0.17	Yb 0.35	Lu 0.16
Th 0.54	Pa	U 0.28	Np 0.06	Pu 0.07	Am 0.10	Cm 0.10	Bk 0.10	Cf 0.10	Es 0.10	Fm 0.10	Md 0.10	No 0.10	Lw 0.10

Fig. A.5. Thermal conductivity of elements in the periodic table.



Molar volume

$cm^3 mol^{-1}$ for solids, liquids

(molar mass)/density for gases, based on liquid density

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H_2** 14.1																	**He** 31.8	K
n=2	**Li** 13.00	**Be** 4.88											**B** 4.68	**C** 5.34	**N_2** 17.3	**O_2** 14.0	**F_2** 17.1	**Ne** 16.8	L
n=3	**Na** 23.70	**Mg** 13.97											**Al** 9.99	**Si** 12.05	**P** 17.0	**S** 15.5	**Cl_2** 18.7	**Ar** 24.2	M
n=4	**K** 45.46	**Ca** 26.02	**Sc** 15.04	**Ti** 10.64	**V** 8.55	**Cr** 7.78	**Mn** 7.35	**Fe** 7.11	**Co** 6.61	**Ni** 6.59	**Cu** 7.11	**Zn** 9.16	**Ga** 11.44	**Ge** 13.57	**As** 13.08	**Se** 16.42	**Br_2** 25.62	**Kr** 32.2	N
n=5	**Rb** 55.79	**Sr** 33.70	**Y** 19.89	**Zr** 14.06	**Nb** 10.84	**Mo** 9.41	**Tc** 8.51	**Ru** 8.22	**Rh** 8.30	**Pd** 8.85	**Ag** 10.27	**Cd** 13.01	**In** 15.73	**Sn** 16.31	**Sb** 18.22	**Te** 20.42	**I_2** 25.74	**Xe** 42.9	O
n=6	**Cs** 70.73	**Ba** 38.21	**La** 22.60	**Hf** 13.41	**Ta** 10.90	**W** 9.50	**Re** 9.07	**Os** 8.41	**Ir** 8.49	**Pt** 9.09	**Au** 10.20	**Hg** 14.81	**Tl** 17.25	**Pb** 18.27	**Bi** 21.37	**Po** 22.73	**At** ---	**Rn** 50.5	P
n=7	**Fr** ---	**Ra** 39.0	**Ac** 22.54	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce 20.70	**Pr** 20.80	**Nd** 20.6	**Pm** 19.95	**Sm** 19.95	**Eu** 28.98	**Gd** 19.90	**Tb** 19.31	**Dy** 19.00	**Ho** 18.74	**Er** 18.45	**Tm** 18.12	**Yb** 24.84	**Lu** 17.76
Th 19.80	**Pa** 15.03	**U** 12.49	**Np** 11.59	**Pu** 12.32	**Am** 17.78	**Cm** 18.29	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

principal quantum number
Highest atomic shell occupied
s-orbital elements
d-orbital elements
p-orbital elements
f-orbital elements

Fig. A.6. Molar volume of elements in the periodic table.

Highest atomic shell occupied

principal quantum number

Specific heat ($Jg^{-1}K^{-1}$)
at 25°C, 1 atm

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	H_2 14.30																	He 5.19	K
n=2	Li 3.58	Be 1.83											B 1.03	C 0.71	N_2 1.04	O_2 0.918	F_2 0.82	Ne 1.03	L
n=3	Na 1.23	Mg 1.02											Al 0.90	Si 0.71	P 0.77	S 0.71	Cl_2 0.48	Ar 0.52	M
n=4	K 0.76	Ca 0.65	Sc 0.568	Ti 0.52	V 0.49	Cr 0.45	Mn 0.48	Fe 0.45	Co 0.42	Ni 0.44	Cu 0.39	Zn 0.39	Ga 0.37	Ge 0.32	As 0.33	Se 0.32	Br_2 0.27	Kr 0.25	N
n=5	Rb 0.36	Sr 0.30	Y 0.30	Zr 0.28	Nb 0.27	Mo 0.25	Tc 0.21	Ru 0.24	Rh 0.24	Pd 0.24	Ag 0.24	Cd 0.23	In 0.23	Sn 0.23	Sb 0.21	Te 0.20	I_2 0.15	Xe 0.16	O
n=6	Cs 0.24	Ba 0.20	La 0.20	Hf 0.14	Ta 0.14	W 0.13	Re 0.14	Os 0.130	Ir 0.13	Pt 0.13	Au 0.13	Hg 0.14	Tl 0.13	Pb 0.13	Bi 0.12	Po ---	At ---	Rn 0.09	P
n=7	Fr ---	Ra ---	Ac 0.12	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

p-orbital elements

d-orbital elements

s-orbital elements

Ce 0.192	Pr 0.193	Nd 0.190	Pm ---	Sm 0.20	Eu 0.18	Gd 0.24	Tb 0.18	Dy 0.17	Ho 0.17	Er 0.17	Tm 0.16	Yb 0.16	Lu 0.15
Th 0.11	Pa ---	U 0.116	Np 0.12	Pu 0.13	Am 0.11	Cm ---	Bk ---	Cf ---	Es ---	Fm ---	Md ---	No ---	Lw ---

f-orbital elements

Fig. A.7. Specific heat of elements in the periodic table.

Atomic radius (0.01 Å)

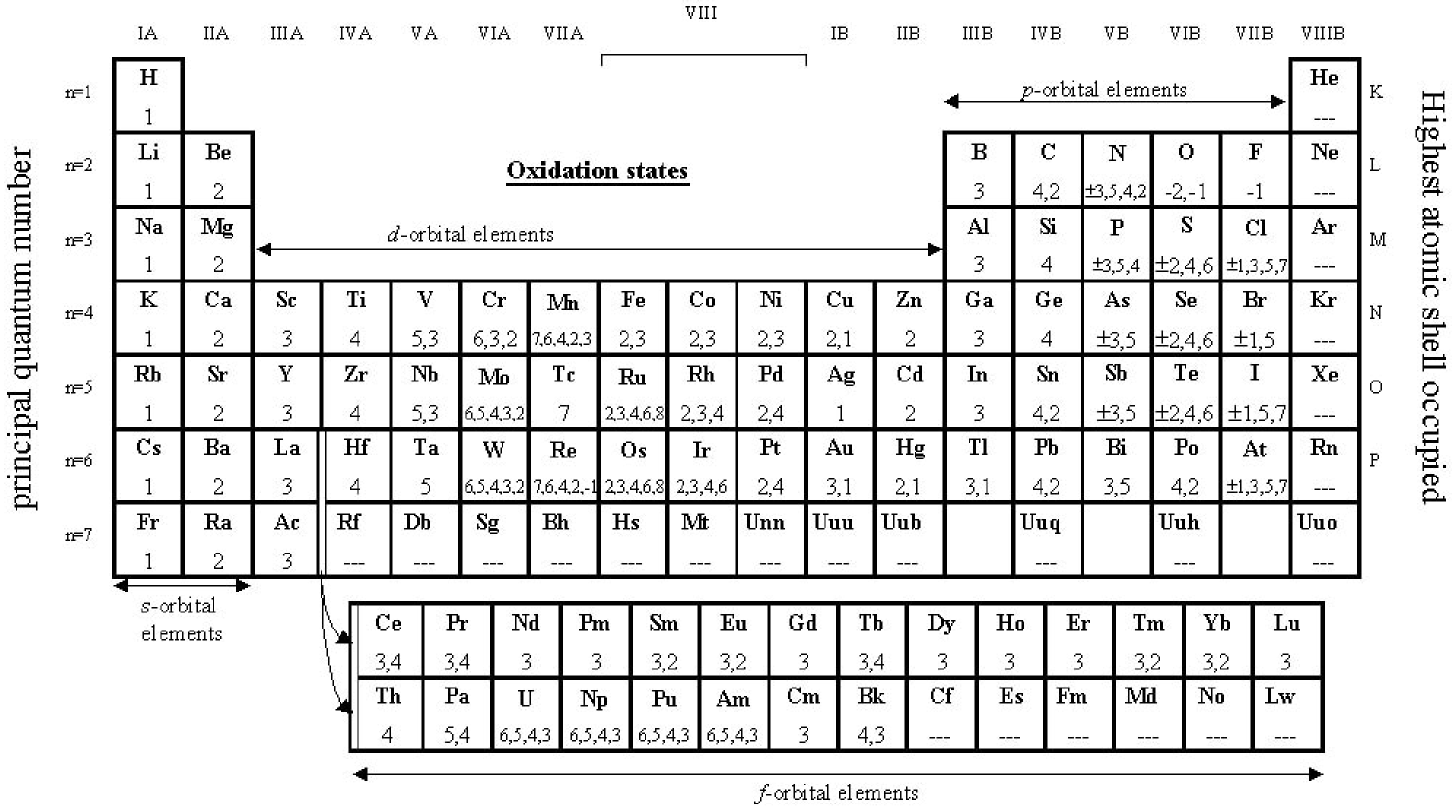
s-orbital elements (IA–IIA); *d*-orbital elements (IIIA–IIB); *p*-orbital elements (IIIB–VIIIB)

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	Highest atomic shell occupied
n=1	H 37																	He ---	K
n=2	Li 152	Be 111											B 80	C 77	N 74	O 74	F 71	Ne ---	L
n=3	Na 186	Mg 160											Al 143	Si 118	P 110	S 103	Cl 99	Ar ---	M
n=4	K 227	Ca 197	Sc 161	Ti 145	V 131	Cr 125	Mn 137	Fe 124	Co 125	Ni 125	Cu 128	Zn 133	Ga 122	Ge 123	As 125	Se 116	Br 114	Kr ---	N
n=5	Rb 248	Sr 215	Y 178	Zr 159	Nb 143	Mo 136	Tc 135	Ru 133	Rh 135	Pd 138	Ag 145	Cd 149	In 163	Sn 141	Sb 145	Te 143	I 133	Xe ---	O
n=6	Cs 265	Ba 217	La 187	Hf 156	Ta 143	W 137	Re 137	Os 134	Ir 136	Pt 139	Au 144	Hg 150	Tl 170	Pb 175	Bi 155	Po 118	At ---	Rn ---	P
n=7	Fr ---	Ra ---	Ac 188	Rf ---	Db ---	Sg ---	Bh ---	Hs ---	Mt ---	Unn ---	Uuu ---	Uub ---		Uuq ---		Uuh ---		Uuo ---	

f-orbital elements

Ce 183	Pr 182	Nd 181	Pm 181	Sm 180	Eu 199	Gd 179	Tb 176	Dy 175	Ho 174	Er 173	Tm 173	Yb 194	Lu 172
Th 180	Pa 161	U 139	Np ---	Pu 151	Am 131	Cm ---	Bk ---	Cf ---	Es ---	Fm ---	Md ---	No ---	Lw ---

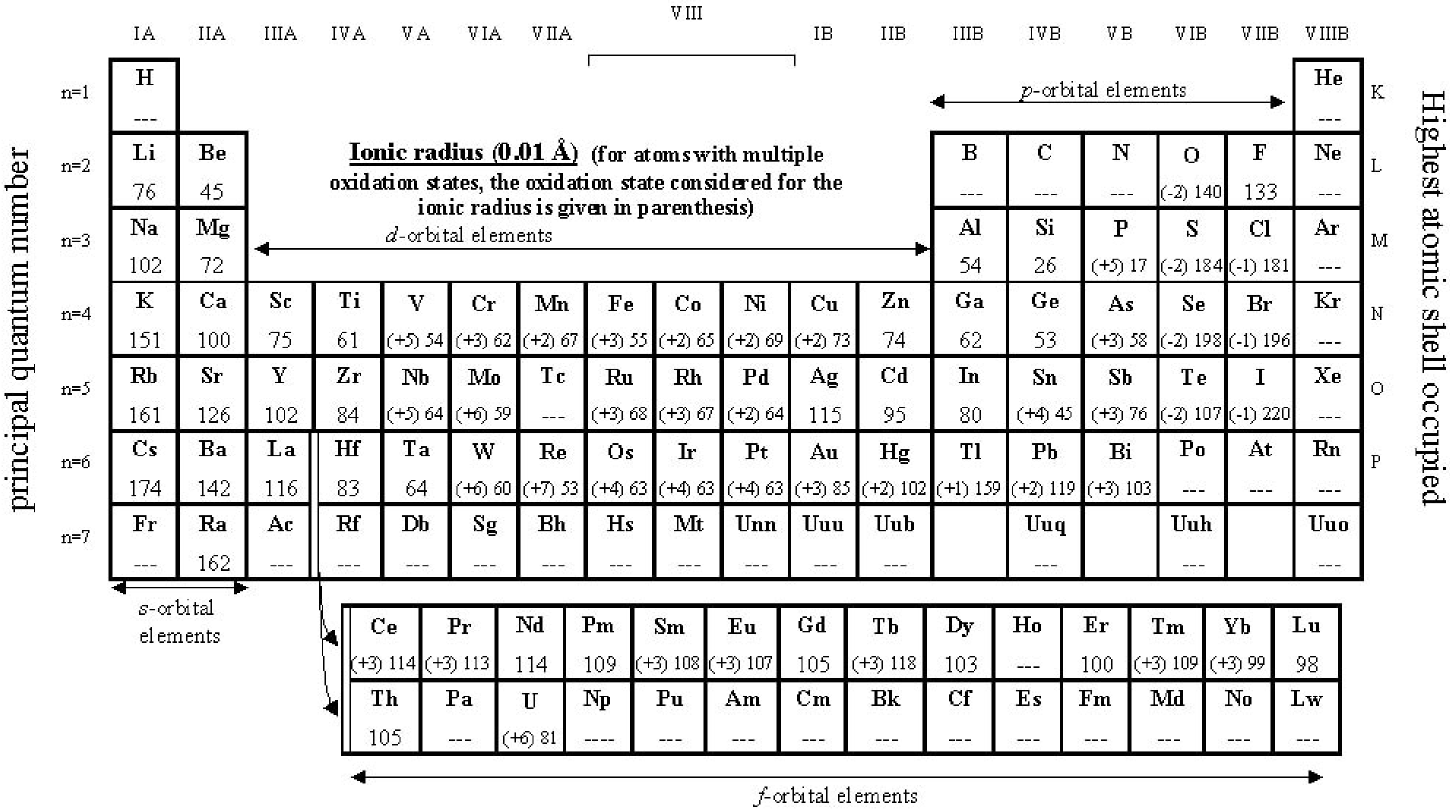
Fig. A.8. Atomic radius of elements in the periodic table.


Highest atomic shell occupied
principal quantum number
Oxidation states
p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H** 1																	**He** ---	K
n=2	**Li** 1	**Be** 2											**B** 3	**C** 4,2	**N** ±3,5,4,2	**O** -2,-1	**F** -1	**Ne** ---	L
n=3	**Na** 1	**Mg** 2											**Al** 3	**Si** 4	**P** ±3,5,4	**S** ±2,4,6	**Cl** ±1,3,5,7	**Ar** ---	M
n=4	**K** 1	**Ca** 2	**Sc** 3	**Ti** 4	**V** 5,3	**Cr** 6,3,2	**Mn** 7,6,4,2,3	**Fe** 2,3	**Co** 2,3	**Ni** 2,3	**Cu** 2,1	**Zn** 2	**Ga** 3	**Ge** 4	**As** ±3,5	**Se** ±2,4,6	**Br** ±1,5	**Kr** ---	N
n=5	**Rb** 1	**Sr** 2	**Y** 3	**Zr** 4	**Nb** 5,3	**Mo** 6,5,4,3,2	**Tc** 7	**Ru** 2,3,4,6,8	**Rh** 2,3,4	**Pd** 2,4	**Ag** 1	**Cd** 2	**In** 3	**Sn** 4,2	**Sb** ±3,5	**Te** ±2,4,6	**I** ±1,5,7	**Xe** ---	O
n=6	**Cs** 1	**Ba** 2	**La** 3	**Hf** 4	**Ta** 5	**W** 6,5,4,3,2	**Re** 7,6,4,2,-1	**Os** 2,3,4,6,8	**Ir** 2,3,4,6	**Pt** 2,4	**Au** 3,1	**Hg** 2,1	**Tl** 3,1	**Pb** 4,2	**Bi** 3,5	**Po** 4,2	**At** ±1,3,5,7	**Rn** ---	P
n=7	**Fr** 1	**Ra** 2	**Ac** 3	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce 3,4	**Pr** 3,4	**Nd** 3	**Pm** 3	**Sm** 3,2	**Eu** 3,2	**Gd** 3	**Tb** 3,4	**Dy** 3	**Ho** 3	**Er** 3	**Tm** 3,2	**Yb** 3,2	**Lu** 3
Th 4	**Pa** 5,4	**U** 6,5,4,3	**Np** 6,5,4,3	**Pu** 6,5,4,3	**Am** 6,5,4,3	**Cm** 3	**Bk** 4,3	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.9. Oxidation states of elements in the periodic table.



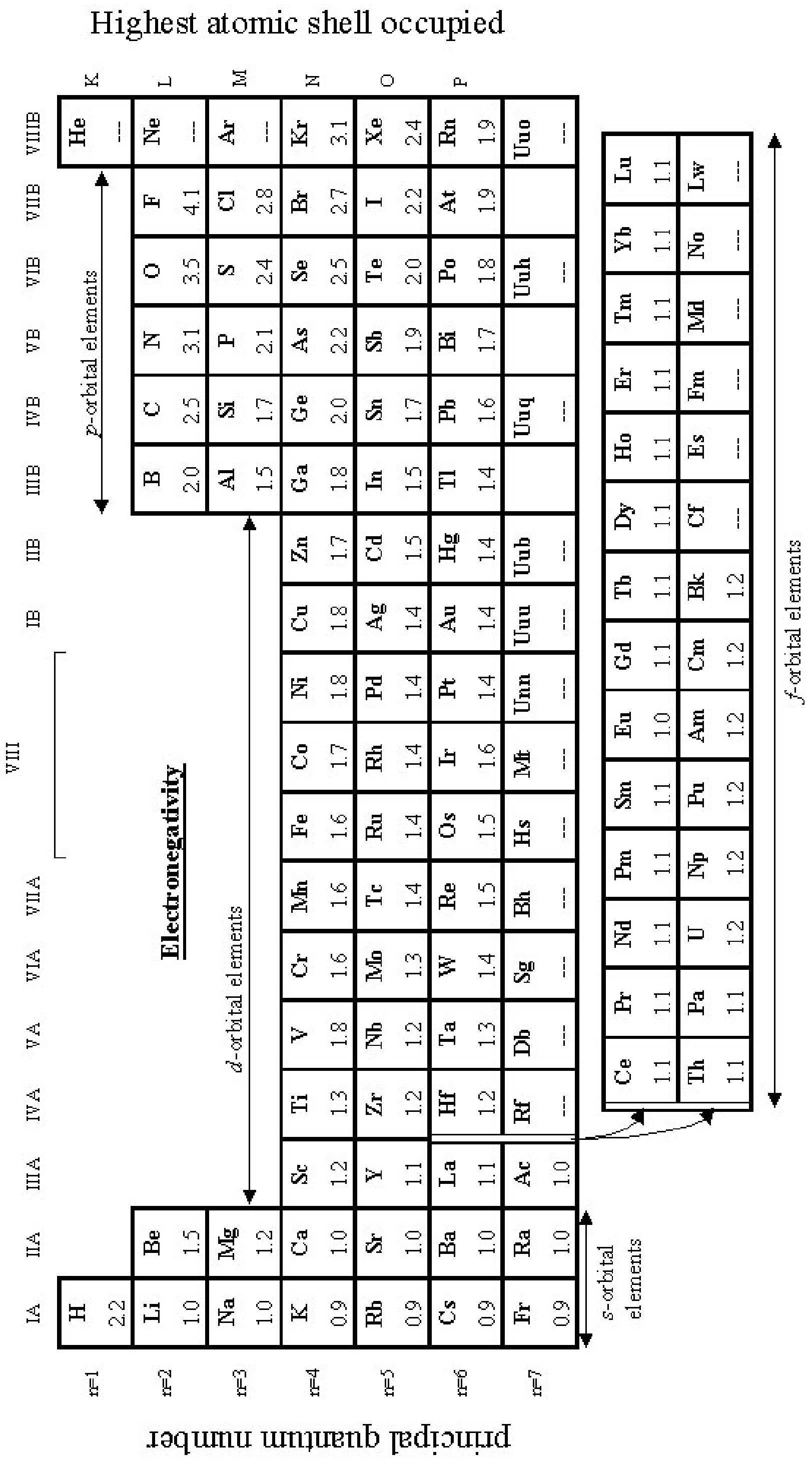
p-orbital elements
d-orbital elements
s-orbital elements
f-orbital elements

Ionic radius (0.01 Å) **(for atoms with multiple oxidation states, the oxidation state considered for the ionic radius is given in parenthesis)**

principal quantum number	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	Highest atomic shell occupied
n=1	**H** ---																	**He** ---	K
n=2	**Li** 76	**Be** 45											**B** ---	**C** ---	**N** ---	**O** (-2) 140	**F** 133	**Ne** ---	L
n=3	**Na** 102	**Mg** 72											**Al** 54	**Si** 26	**P** (+5) 17	**S** (-2) 184	**Cl** (-1) 181	**Ar** ---	M
n=4	**K** 151	**Ca** 100	**Sc** 75	**Ti** 61	**V** (+5) 54	**Cr** (+3) 62	**Mn** (+2) 67	**Fe** (+3) 55	**Co** (+2) 65	**Ni** (+2) 69	**Cu** (+2) 73	**Zn** 74	**Ga** 62	**Ge** 53	**As** (+3) 58	**Se** (-2) 198	**Br** (-1) 196	**Kr** ---	N
n=5	**Rb** 161	**Sr** 126	**Y** 102	**Zr** 84	**Nb** (+5) 64	**Mo** (+6) 59	**Tc** ---	**Ru** (+3) 68	**Rh** (+3) 67	**Pd** (+2) 64	**Ag** 115	**Cd** 95	**In** 80	**Sn** (+4) 45	**Sb** (+3) 76	**Te** (-2) 107	**I** (-1) 220	**Xe** ---	O
n=6	**Cs** 174	**Ba** 142	**La** 116	**Hf** 83	**Ta** 64	**W** (+6) 60	**Re** (+7) 53	**Os** (+4) 63	**Ir** (+4) 63	**Pt** (+4) 63	**Au** (+3) 85	**Hg** (+2) 102	**Tl** (+1) 159	**Pb** (+2) 119	**Bi** (+3) 103	**Po** ---	**At** ---	**Rn** ---	P
n=7	**Fr** ---	**Ra** 162	**Ac** ---	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce (+3) 114	**Pr** (+3) 113	**Nd** 114	**Pm** 109	**Sm** (+3) 108	**Eu** (+3) 107	**Gd** 105	**Tb** (+3) 118	**Dy** 103	**Ho** ---	**Er** 100	**Tm** (+3) 109	**Yb** (+3) 99	**Lu** 98
Th 105	**Pa** ---	**U** (+6) 81	**Np** ----	**Pu** ---	**Am** ---	**Cm** ---	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.10. Ionic radius of elements in the periodic table.

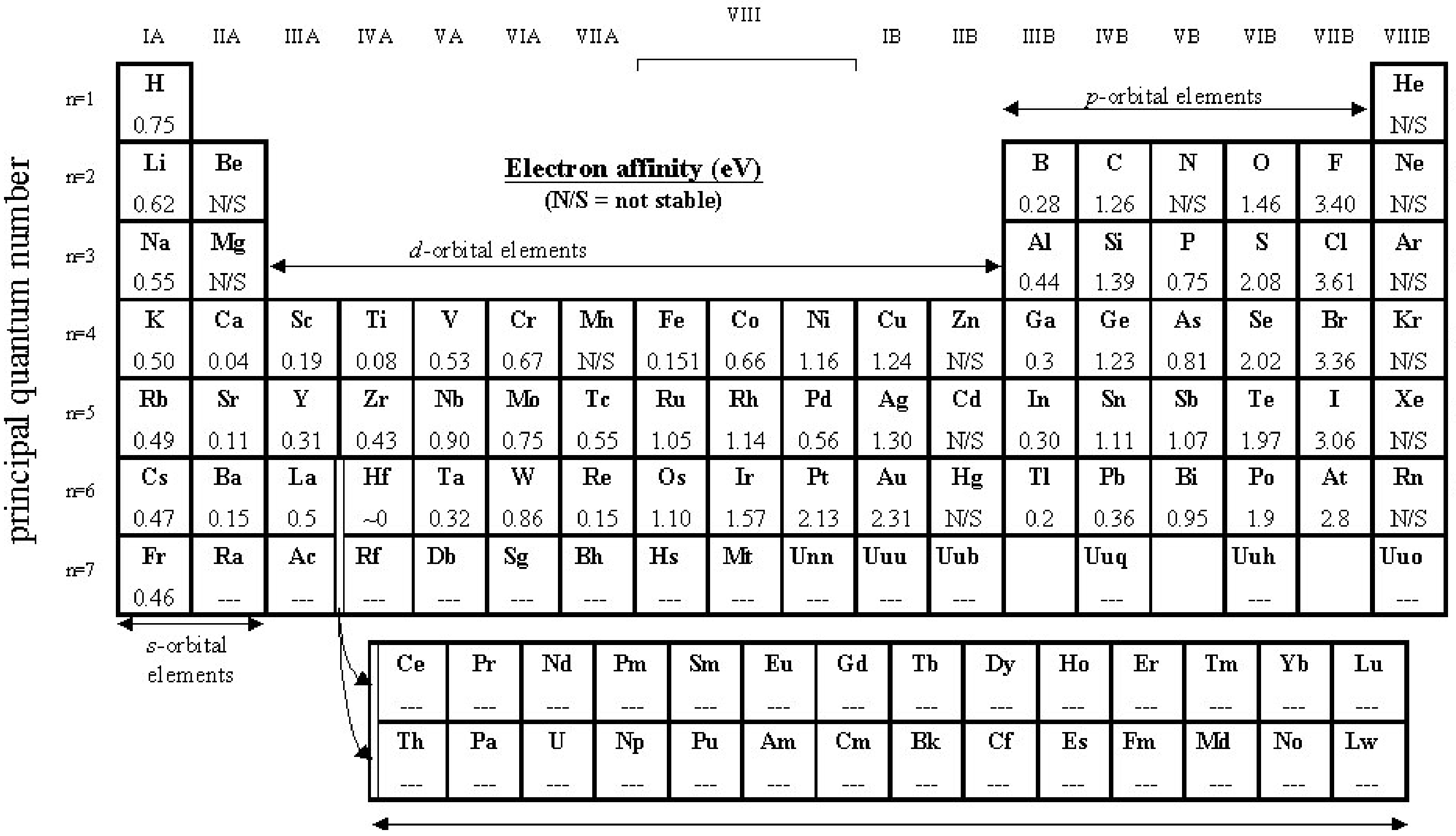

Highest atomic shell occupied
principal quantum number
s-orbital elements
d-orbital elements
p-orbital elements
f-orbital elements

Electronegativity

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB	
n=1	**H** 2.2																	**He** ---	K
n=2	**Li** 1.0	**Be** 1.5											**B** 2.0	**C** 2.5	**N** 3.1	**O** 3.5	**F** 4.1	**Ne** ---	L
n=3	**Na** 1.0	**Mg** 1.2											**Al** 1.5	**Si** 1.7	**P** 2.1	**S** 2.4	**Cl** 2.8	**Ar** ---	M
n=4	**K** 0.9	**Ca** 1.0	**Sc** 1.2	**Ti** 1.3	**V** 1.8	**Cr** 1.6	**Mn** 1.6	**Fe** 1.6	**Co** 1.7	**Ni** 1.8	**Cu** 1.8	**Zn** 1.7	**Ga** 1.8	**Ge** 2.0	**As** 2.2	**Se** 2.5	**Br** 2.7	**Kr** 3.1	N
n=5	**Rb** 0.9	**Sr** 1.0	**Y** 1.1	**Zr** 1.2	**Nb** 1.2	**Mo** 1.3	**Tc** 1.4	**Ru** 1.4	**Rh** 1.4	**Pd** 1.4	**Ag** 1.4	**Cd** 1.5	**In** 1.5	**Sn** 1.7	**Sb** 1.9	**Te** 2.0	**I** 2.2	**Xe** 2.4	O
n=6	**Cs** 0.9	**Ba** 1.0	**La** 1.1	**Hf** 1.2	**Ta** 1.3	**W** 1.4	**Re** 1.5	**Os** 1.5	**Ir** 1.6	**Pt** 1.4	**Au** 1.4	**Hg** 1.4	**Tl** 1.4	**Pb** 1.6	**Bi** 1.7	**Po** 1.8	**At** 1.9	**Rn** 1.9	P
n=7	**Fr** 0.9	**Ra** 1.0	**Ac** 1.0	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---	

Ce 1.1	**Pr** 1.1	**Nd** 1.1	**Pm** 1.1	**Sm** 1.1	**Eu** 1.0	**Gd** 1.1	**Tb** 1.1	**Dy** 1.1	**Ho** 1.1	**Er** 1.1	**Tm** 1.1	**Yb** 1.1	**Lu** 1.1
Th 1.1	**Pa** 1.1	**U** 1.2	**Np** 1.2	**Pu** 1.2	**Am** 1.2	**Cm** 1.2	**Bk** 1.2	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.11. Electronegativity of elements in the periodic table.



principal quantum number
Electron affinity (eV)
(N/S = not stable)
p-orbital elements
d-orbital elements
s-orbital elements

	IA	IIA	IIIA	IVA	VA	VIA	VIIA	VIII			IB	IIB	IIIB	IVB	VB	VIB	VIIB	VIIIB
n=1	**H** 0.75																	**He** N/S
n=2	**Li** 0.62	**Be** N/S											**B** 0.28	**C** 1.26	**N** N/S	**O** 1.46	**F** 3.40	**Ne** N/S
n=3	**Na** 0.55	**Mg** N/S											**Al** 0.44	**Si** 1.39	**P** 0.75	**S** 2.08	**Cl** 3.61	**Ar** N/S
n=4	**K** 0.50	**Ca** 0.04	**Sc** 0.19	**Ti** 0.08	**V** 0.53	**Cr** 0.67	**Mn** N/S	**Fe** 0.151	**Co** 0.66	**Ni** 1.16	**Cu** 1.24	**Zn** N/S	**Ga** 0.3	**Ge** 1.23	**As** 0.81	**Se** 2.02	**Br** 3.36	**Kr** N/S
n=5	**Rb** 0.49	**Sr** 0.11	**Y** 0.31	**Zr** 0.43	**Nb** 0.90	**Mo** 0.75	**Tc** 0.55	**Ru** 1.05	**Rh** 1.14	**Pd** 0.56	**Ag** 1.30	**Cd** N/S	**In** 0.30	**Sn** 1.11	**Sb** 1.07	**Te** 1.97	**I** 3.06	**Xe** N/S
n=6	**Cs** 0.47	**Ba** 0.15	**La** 0.5	**Hf** ~0	**Ta** 0.32	**W** 0.86	**Re** 0.15	**Os** 1.10	**Ir** 1.57	**Pt** 2.13	**Au** 2.31	**Hg** N/S	**Tl** 0.2	**Pb** 0.36	**Bi** 0.95	**Po** 1.9	**At** 2.8	**Rn** N/S
n=7	**Fr** 0.46	**Ra** ---	**Ac** ---	**Rf** ---	**Db** ---	**Sg** ---	**Bh** ---	**Hs** ---	**Mt** ---	**Unn** ---	**Uuu** ---	**Uub** ---		**Uuq** ---		**Uuh** ---		**Uuo** ---

Ce ---	**Pr** ---	**Nd** ---	**Pm** ---	**Sm** ---	**Eu** ---	**Gd** ---	**Tb** ---	**Dy** ---	**Ho** ---	**Er** ---	**Tm** ---	**Yb** ---	**Lu** ---
Th ---	**Pa** ---	**U** ---	**Np** ---	**Pu** ---	**Am** ---	**Cm** ---	**Bk** ---	**Cf** ---	**Es** ---	**Fm** ---	**Md** ---	**No** ---	**Lw** ---

Fig. A.12. Electron affinity of elements in the periodic table.

A.4. Physical properties of important semiconductors

Semiconductor		Bandgap energy (eV) 300 K	 0 K	Band	ε
Element	C	5.47	5.48	indirect	5.7
	Si	1.12	1.17	indirect	11.9
	Ge	0.66	0.74	indirect	16.0
	Sn		0.082	direct	
IV-IV	α-SiC	2.996	3.03	indirect	10.0
III-V	BN	~7.5		indirect	7.1
	GaN	3.36	3.50	direct	12.2
	GaP	2.26	2.34	indirect	11.1
	BP	2.0			
	AlSb	1.58	1.68	indirect	14.4
	GaAs	1.42	1.52	direct	13.1
	InP	1.35	1.42	direct	12.4
	GaSb	0.72	0.81	direct	15.7
	InAs	0.36	0.42	direct	14.6
	InSb	0.17	0.23	direct	17.7
II-VI	ZnS	3.68	3.84	direct	5.2
	ZnO	3.35	3.42	direct	9.0
	CdS	2.42	2.56	direct	5.4
	CdSe	1.70	1.85	direct	10.0
	CdTe	1.56		direct	10.2
IV-VI	PbS	0.41	0.286	indirect	17.0
	PbTe	0.31	0.19	indirect	30.0

Semiconductor		Mobility at 300 K ($cm^2/V \cdot s$)		Effective masses (in units of m_0)	
		electrons	holes	electrons m_e	holes m_h
Element	C	1800	1200	0.2	0.25
	Si	1500	450	0.98[a] 0.19[b]	0.16[c] 0.49[d]
	Ge	3900	1900	1.64[a] 0.082[b]	0.04[c] 0.28[d]
	Sn	1400	1200		
IV-IV	α-SiC	400	50	0.60	1.00
III-V	BN				
	GaN	380		0.19	0.60
	GaP	100	75	0.82	0.60
	BP				
	AlSb	200	420	0.12	0.98
	GaAs	8500	400	0.067	0.082[c] 0.45[d]
	InP	4600	150	0.077	0.64
	GaSb	5000	850	0.42	0.04[c] 0.4[d]
	InAs	33000	460	0.023	0.40
	InSb	80000	1250	0.0145	0.40
II-VI	ZnS	165	5	0.40	
	ZnO	200	180	0.27	
	CdS	340	50	0.21	0.80
	CdSe	800		0.13	0.45
	CdTe	1050	100		
IV-VI	PbS	600	700	0.25	0.25
	PbTe	6000	4000	0.17	0.20

[a] Longitudinal effective mass.

[b] Transverse effective mass.

[c] Light-hole effective mass.

[d] Heavy-hole effective mass.

Semiconductor	Intrinsic carrier concentration at 300 K (cm^{-3})
Ge	2.4×10^{13}
Si	1.45×10^{10}
GaAs	2.15×10^{6}

References

Bethe, H.A., “Theory of the boundary layer of crystal rectifiers,” *MIT Radiation Laboratory Report* 43-12, 1942.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

A.5. Thermionic emission

The thermionic emission theory is a semi-classical approach developed by Bethe [1942], which accurately describes the transport of electrons through a semiconductor-metal junction. The parameters taken into account are the temperature T, the energy barrier height $q\Phi_B$ and the bias voltage V between the far-ends of the semiconductor and the metal. These quantities are illustrated in Fig. A.13.

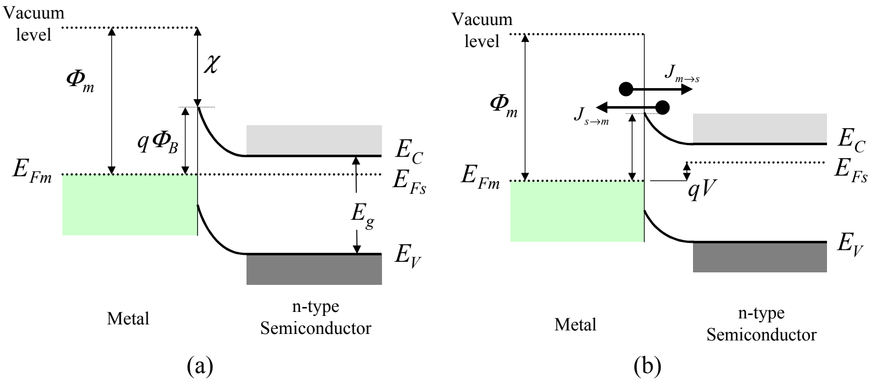


Fig. A.13. Energy band diagram of a Schottky metal-(n-type) semiconductor junction: (a) at equilibrium and (b) under forward bias (V>0), showing the transport of electrons over the potential barrier as the main transport process under forward bias.

The theory is based on the following three assumptions: (i) the energy barrier height $q\Phi_B$ at the interface is much higher than k_bT, (ii) the junction plane is at thermal equilibrium, (iii) this equilibrium is not affected by the presence of an electrical current. By assuming these, the thermionic emission current only depends on the energy barrier height and not its spatial profile. Furthermore, the total current is therefore the sum of the current from the semiconductor into the metal, denoted $J_{s\to m}$, and that of the metal into the semiconductor, denoted $J_{m\to s}$.

To calculate the first current, $J_{s\to m}$, the theory assumes that the energy of the electrons in the conduction band is purely kinetic, and that their velocity is distributed isotropically. The current density from the semiconductor into the metal can be calculated by summing the current contribution from all the electrons that have an energy higher than the barrier $q\Phi_B$ and that have a velocity component from the semiconductor toward the metal. This results in the following expression:

Eq. (A.1) $$J_{s\to m} = \left(\frac{4\pi q m^* k_b^2}{h^3}\right) T^2 e^{-\frac{q\Phi_B}{k_b T}} e^{\frac{qV}{k_b T}}$$

or:

Eq. (A.2) $$J_{s\to m} = A^* T^2 e^{-\frac{q\Phi_B}{k_b T}} e^{\frac{qV}{k_b T}}$$

where k_b is the Boltzmann constant, V is the bias voltage, Φ_B is the barrier height, T is the temperature in degrees Kelvin, h is Plank's constant and m^* is the electron effective mass in the direction perpendicular to the junction plane, and $A^* = \frac{4\pi q m^* k_b^2}{h^3}$ is called the effective Richardson constant for thermionic emission. This quantity can be related to the Richardson constant for free electrons, $A = 120\ \mathrm{A\cdot cm^{-2}\cdot K^{-2}}$, as discussed below.

For n-type semiconductors with an isotropic electron effective mass m^* in the minimum of the conduction band, we have $\frac{A^*}{A} = \frac{m^*}{m_0}$, where m_0 is the electron rest mass.

For n-type semiconductors with a multiple-valley conduction band, the effective Richardson constant A^* associated with each local energy minimum is given by $\frac{A^*}{A} = \frac{\left(l_x^2 m_y^* m_z^* + l_y^2 m_z^* m_x^* + l_z^2 m_x^* m_y^*\right)^{\frac{1}{2}}}{m_0}$, where l_x, l_y and l_z are the direction cosines corresponding to this energy minimum in the First Brillouin zone.

In the case of a p-type semiconductor, we need to consider the heavy-hole and the light-hole bands in the valence band, both of which have their maximum at the center of the Brillouin zone. The effective Richardson constant is then given by the following expression $\frac{A^*}{A} = \frac{\left(m_{lh}^* + m_{hh}^*\right)}{m_0}$, where m_{hh}^* and m_{lh}^* are the heavy-hole and light-hole effective masses, respectively. A few examples of values for $\frac{A^*}{A}$ are given in Table A.1.

Semiconductor	Si	Ge	GaAs
n-type <111>	2.2	1.11	0.068 (low field) 1.2 (high field)
n-type <100>	2.1	1.19	0.068 (low field) 1.2 (high field)
p-type	0.66	0.34	0.62

Table A.1. Examples of values for $\frac{A^*}{A}$ *in a few semiconductors. [Sze 1981]*

The second current contribution to the thermionic emission current is the current flowing from the metal into the semiconductor, $J_{m\to s}$. As the barrier height for the transport of electrons in this direction is independent of the applied bias voltage V (Fig. A.13(b)), $J_{m\to s}$ is also independent of the bias voltage. $J_{m\to s}$ is therefore equal to the opposite of $J_{s\to m}$ when $V = 0$, because no net current exists at equilibrium. Using Eq. (A.2), we obtain:

Eq. (A.3)
$$J_{m\to s} = -A^* T^2 e^{-\frac{q\Phi_B}{k_b T}}$$

The total current density is therefore:

Eq. (A.4)
$$\begin{aligned} J = J_{s\to m} + J_{m\to s} &= A^* T^2 e^{-\frac{q\Phi_B}{k_b T}} \left[e^{\frac{qV}{k_b T}} - 1 \right] \\ &= J_{ST} \left[e^{\frac{qV}{k_b T}} - 1 \right] \end{aligned}$$

This expression shows that the thermionic emission current resembles the diode equation obtained in Eq. (4.56). The difference lies in the saturation current density which is now given by:

Eq. (A.5) $$J_{ST} = A^{*}T^{2}\, e^{-\frac{q\Phi_B}{k_b T}}$$

References

Bethe, H.A., “Theory of the boundary layer of crystal rectifiers,” *MIT Radiation Laboratory Report* 43-12, 1942.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

A.6. Minority carrier lifetime measurement

Significance of carrier lifetime

A simple definition of minority carrier lifetime is the amount of time that a minority carrier in a semiconductor can survive before recombining with a majority carrier. There are three main mechanisms for the electron-hole recombination in a semiconductor. If the electron and hole recombine and give their energy to a photon, it is called radiative recombination. If they recombine and give their energy to a third electron or hole it is called Auger recombination and if they recombine through mid-gap traps, it is called Shockley-Read-Hall recombination. The effective lifetime of a carrier is the combination of these three, as shown in Eq. (A.6):

Eq. (A.6) $$\frac{1}{\tau_{eff}} = \frac{1}{\tau_{SRH}} + \frac{1}{\tau_{Auger}} + \frac{1}{\tau_{Rad}}$$

In a photodiode, photo-generated carriers from the region that is more than one diffusion length away from the depletion region will not contribute to the photo-current since carriers recombine before they reach the junction edge. This has been shown schematically in Fig. A.14. The diffusion length is related to carrier lifetime as followed:

Eq. (A.7) $$L = \sqrt{D\tau}$$

Where L is the diffusion length, τ is the carrier lifetime, and D is the diffusivity given by the Einstein relation:

Eq. (A.8) $$D = \frac{\mu k_B T}{q}$$

Where μ is the mobility, k_B is the Boltzmann constant, T is the temperature and q is the electron charge. It is clear that a material with longer carrier lifetime would have longer diffusion length, and thus better optical efficiency.

Fig. A.14. The carriers generated within a diffusion length of the junction are collected while the rest are lost

The dark current in a photodiode also depends on carrier lifetime. As shown in Eq. (A.9), the dark current consists of four major components: diffusion, generation-recombination, band to band tunneling and trap-assisted tunneling.

Eq. (A.9) $$J_{dark} = J_{diff} + J_{G-R} + J_{BTB} + J_{TAT}$$

The diffusion and generation-recombination components of the dark current strongly depend on carrier lifetime, as shown in Eq. (A.10) and Eq. (A.11)

Eq. (A.10)

Eq. (A.11)

for both of these equations the lifetime appears in denominator which means that for achieving low dark currents, long lifetimes are desirable.

Different carrier lifetime measurement techniques

Carrier lifetime measurement techniques fall into two main categories: they either need to have electrical contacts or they do not need electrical contacts and can be done at the wafer level. From the contact methods we can name photodiode modeling, photoconductor modeling, reverse recovery, EBIC (Electron Beam Induced Current) and photo-electromagnetic effects. There are two methods that can measure the carrier lifetime at the wafer level: photoconductive decay and time-resolved photoluminescence.

Photodiode modeling

Eq. (A.10) shows the relationship between the diffusion component of a photodiode dark current and the carrier lifetime. If we measure the dark current by changing different parameters in Eq. (A.10) when we are in diffusion limited regime, we can extract the carrier lifetime. In Fig. A.15 you can see the experimental extraction of carrier lifetime from the measurement of Type-II photodetectors, as were discussed in Chapter 10.

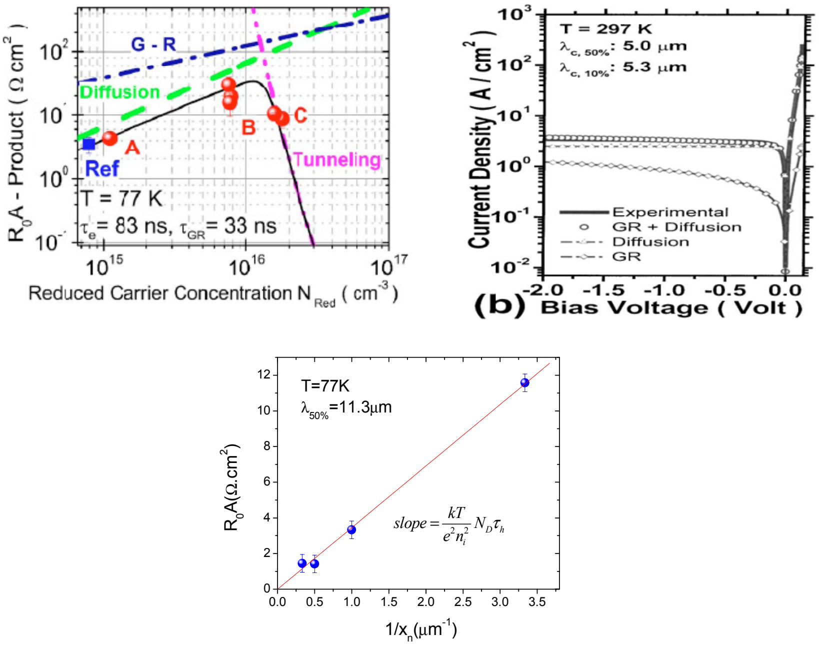


Fig. A.15. Extraction of carrier lifetime based on I-V modeling of photodiode a) Diffusion R_0A versus carrier concentration b) Dark current vs. voltage c) Diffusion R_0A vs. device thickness

Photoconductor modeling

The effective lifetime of the carriers can also be extracted from the optical and electrical measurements of photoconductors with a geometry similar to that shown in Fig. A.16 by using the expression:

Eq. (A.12) $$\tau_{effective} = \frac{lg}{E(\mu_e + \mu_p)}$$

where l is the device length, g is the photoconductor gain, and E is the electrical field. The gain of the device can be calculated from g €$R_i hc / \lambda\eta q$ where R is the current responsivity. Hall measurement can be used to measure the mobilities. Assuming an internal quantum efficiency near unity and negligible reflection from the bottom of the detector and unpolished backside of the substrate, the quantum efficiency can be calculated. Given these experimentally determined values, a carrier lifetime can thus be calculated. [Mohseni et al. 1998] [Mohseni et al. 1999]

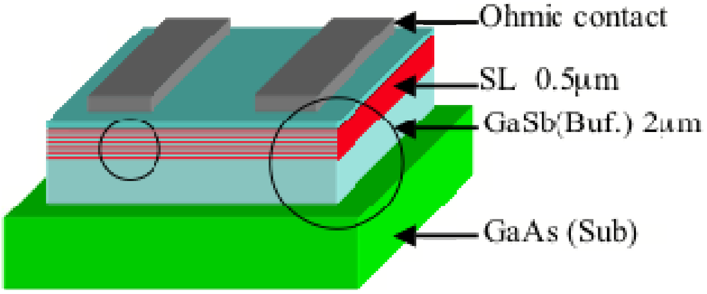


Fig. A.16. Schematic diagram of a Type – II InAs/GaSb Superlattice photoconductor used to extracted carrier lifetime.

Disadvantages of measurement techniques using electrical contacts

One of the reasons to pursue carrier lifetime measurement is to investigate the material at the wafer level. This can provide a fast feedback for the material quality. It can also be used to systematically study the material uniformity before processing devices. However, measurements that require electrical contacts cannot meet this need for several reasons:

- The processing required to prepare the samples might change the carrier dynamics by introducing surface recombinations and it can obscure the wafer quality.
- There is no way to extract different mechanisms of recombination(SRH, Auger, and radiative).
- The uniformity of the wafers cannot be studied because once the sample is processed, the non-uniformity can be caused by the processing.
- The extraction is based on the parameters that are not well known and cannot be directly measured, such as mobility.

For the above reasons, there is interest in developing measurement techniques that can be done on the wafer level.

Non-contact photoconductive decay

In Fig. A.17, a schematic diagram of contact-less photoconductive decay measurement system is shown. Through optical excitation electro-hole pairs are created., then by removing the excitation, the signal decay from the material is monitored by recording the reflection or transmission of the microwave or RF signal. This is a very simple technique at room temperature, but is not well adapted to cryogenic measurements within the confines of a cryostat. It also has difficulty measuring lifetimes lower than 100 ns. [Schroder 2006]

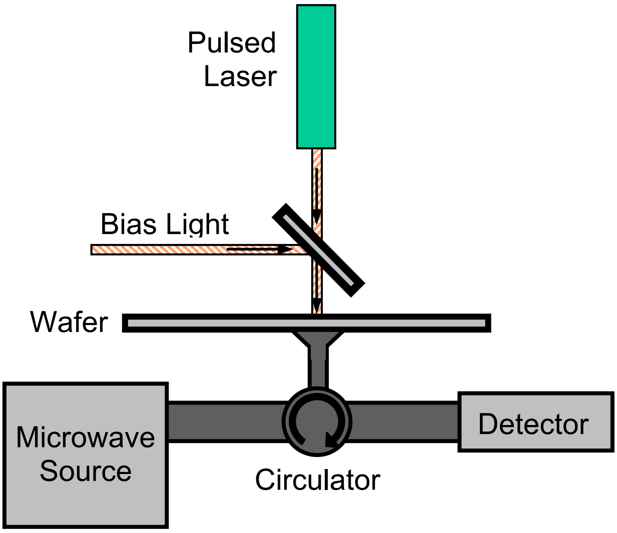


Fig. A.17. Schematic diagram of a microwave-reflection based photoconductive decay measurement system.

Time-resolved photoluminescence

In time-resolved photoluminescence, the excess carrier generation is done by a short pulse of photons having an energy greater than the semiconductor bandgap. The excess carrier density, is then monitored as a function of the time variation of the light emission by the recombining electron-hole pairs that are emitted out of the volume of the semiconductor. As long as the material has a direct bandgap, and has an efficient emission process, this process can be very accurate. As well, by adjusting the laser power and wavelength intensity the penetration depth of the photons can go beyond the surface and reveal the bulk lifetime. Although in principle, the time-resolved

PL setup is quite straightforward the use of such a system for the MWIR to the VLWIR is rather complex due to the pump requirements. High-speed lasers that have pulse widths shorter than the expected lifetime are required. Since we expect the lifetime to be in nanosecond range, the pulse width of the laser must be shorter. A MWIR-VLWIR laser with such characteristics does not exist, so an Optical Parametric Amplifier (OPA) is needed in conjunction with a common pulsed laser to tune the output of the laser. As well, the collection detector must also be able to have a fast respond to low-level signal to reduce noise and get accurate measurements. The response time of the detector needs to be shorter than the lifetime being measured. This is also a challenge since fast detectors commercially available in the wavelengths of interest have either very slow response or low detectivity. Several solutions have been proposed for solving this problems [Shah 1988] [McCahon et al. 1996] [Kost et al. 2005]. But all of them increase the complexity and cost of the setup significantly. This is in contradiction with the main goal to provide a fast and simple feedback of the material.

One of the main advantages of time-resolved photoluminescence carrier lifetime measurement system, besides being a direct and fast way of measuring the carrier lifetime, is that with a proper analysis of the data, it can potentially help in determining the different components of the lifetime. Eq. (A.13) shows the dependence of the recombination rate on the excess carrier density:

Eq. (A.13) $$-\frac{dn}{dt} = A_{SRH}.n + B_{Rad}.n^2 + C_{Auger}.n^3$$

The laser power can be changed to provide different excess carrier densities and a proper fitting of Eq. (A.13)will give us different coefficients. This information would be invaluable in terms of providing an insight into the physics of our material and the ability to tailor it in a way that has never been possible before.

References

Kost, A.R., Pfeiffer, U., and Döhler, G., *Superlattices and Microstructures* Vol. 37, "Time-resolved photocarrier decay for mid-infrared semiconductors with excitation correlation," pp. 373-379, 2005.

McCahon, S.W., Anson, S.A., Jang, D.J., Flatte, M.E., Boggess, T.F., Chow, D.H., Hasenberg, T.C., and Grein, C.H., *Applied Physics Letters* Vol. 68,"Carrier recombination dynamics in a (GaInSb/InAs)/AlGaSb superlattice multiple quantum well," pp. 2135-2137, 1996.Mohseni, H., Litvinov, V.I., and Razeghi, M. *Physical Review B* Vol. 58, "Interface-induced suppression of the Auger recombination in type-II InAs/GaSb superlattices," pp. 15378, 1998.

Mohseni, H., Wojkowski, J., Razeghi, M., Brown, G., and Mitchel, W., *IEEE Journal of Quantum Electronics* Vol. 35, "Uncooled InAs-GaSb type-II infrared detectors grown on GaAs substrates for the 8–12 µm atmospheric window," pp. 1041-1044, 1999.

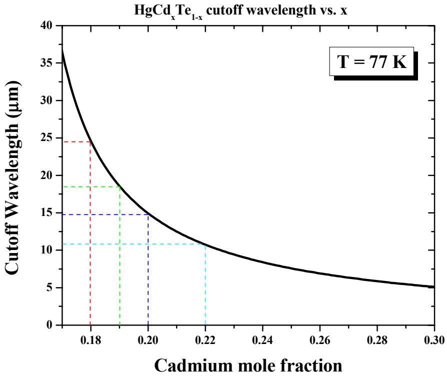
Schroder, D., *Semiconductor Material and Device Characterization*, pp.401, John Wiley & Sons, 2006.

Shah, J., *IEEE Journal of Quantum Electronics* Vol. 24, "Ultrafast luminescence spectroscopy using sum frequency generation," pp. 276-288, 1988.

A.7. Advanced topics in Type-II photodetectors

Introduction

High performance focal plane arrays (FPAs) in the long and very long wavelength infrared (LWIR & VLWIR) spectral bands are highly desirable. Commercially available infrared FPAs at these wavelengths include Quantum Well Photo-detectors (QWIP) and Mercury Cadmium Telluride (HgCdTe or MCT) compounds. However for wavelengths longer than 10 µm, the dark current level of QWIP would significantly limit the FPA operating temperature [Chu 2005]. The current state-of-the-art VLWIR detection technology is thus based on MCT, which can achieve excellent performance in term of sensitivity and speed between 9 and 12 µm [Gilmore 2005] [Manissadjian 2005].



Sensitivity of E_g on Cd%

Cadmium Mole Fraction (x)	Cutoff Wavelength (µm)	λ_c for -2% Δx (µm)
.18	24.60	28.27
.19	18.57	20.59
.20	14.93	16.2
.22	10.7	11.6

Fig. A.18. Left: Variation of the cutoff wavelength with the cadmium concentration at 77K. Right: Effect of cadmium concentration on the uniformity of the cutoff wavelength.

MCT suffers from a number of major drawbacks. The required CdTe substrates are very expensive. The spatial uniformity of MCT is very poor at the composition needed for VLWIR detectors. Very tight control of the cadmium mole fraction must be maintained in order to have a uniform cutoff wavelength and thus a uniform pixel response across the wafer (Fig. A.18). The operability of VLWIR arrays (~80%) is thus limited by the wide variation of the optical response and noise due to the non-uniform bandgap of the material. The limited yield of this technology drives the cost of focal plane arrays prohibitively high. In addition, due to the electronic band structure of MCT, the minority carrier lifetime strongly degrades at elevated temperatures as a result of non-radiative Auger recombination. Thus, in order to enable higher temperature operation using MCT, one has to compromise on the sensitivity of the detectors.

Advantages of Type-II

Type-II InAs/GaSb superlattices represent the most promising material system capable of delivering more affordable and producible FPAs than the current technologies, while at the same time exhibiting similar or better performance. Owing to its coupled quantum well-based design, the cutoff wavelength across a typical 2" wafer is relatively insensitive to normal variations in layer thicknesses, and thus uniform infrared materials can be grown. Not only does this have ramifications for device performance but from a cost and yield standpoint the advantages are very clear. Additionally, the band structures of these multilayers can be engineered to suppress Auger recombination mechanisms for certain layer compositions, with a wide wavelength coverage range from 2 μm up to at least 32 μm [Youngdale 1994] [Zegrya 1995] [Mohseni 1998]. Experimental results have shown a nearly one order of magnitude lower Auger recombination rate at room temperature in such detectors compared to typical HgCdTe detectors with similar bandgap [Mohseni 2000]. Furthermore, the larger electron effective mass can lead to a much reduced dark current. The strong bonding between group III and group V elements leads to very stable materials and high uniformity.

The main advantage of Type-II InAs/GaSb superlattices is the possibility to adjust the position of the conduction and valence band of the material. The bottom of the conduction band can easily be modified by increasing or decreasing the thickness of the InAs and GaSb layers. However, it is necessary to add an AlSb layer inside the GaSb to modify the position of the valence band. This has been shown to allow the possibility to independently tune the position of the different bands. These new devices can potentially improve the sensitivity and uniformity of VLWIR FPAs.

In the interest of realizing high performance type-II devices a number of novel structures and designs have been proposed over the years. This appendix presents some of the most promising technological innovation in type-II.

Nano-Planar Junction (NPJ)

The new concept discussed here consists in fabricating a Nano-Planar Junction (NPJ) using type-II superlattices. The NPJ is realized by locally doping *n*-type (or *p*-type) a π-type (or ν-type) semiconductors (Fig. A.19).

The infrared photons are absorbed inside the slightly *p*-doped material, creating electron-hole pairs. As both the lateral (40 µm) and vertical (5–15 µm) diffusion length of carriers are high, the electron and hole diffuse to the *p-n* junction and are separated by the electric field. The carriers are then collected through the top and substrate contacts.

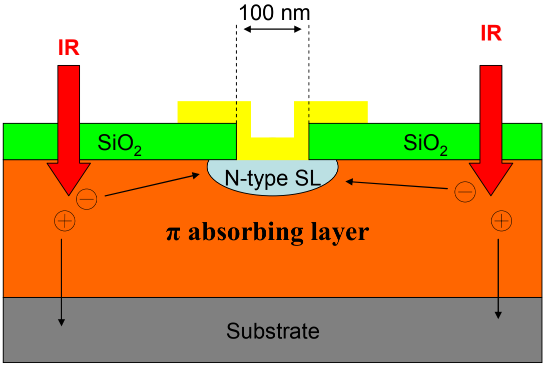


Fig. A.19. Sketch of the nano-planar junction (NPJ).

Advantages of NPJ

The dark current of a *p-n* junction is the result of four main components:
Diffusion current:

Eq. (A.14) $$I_D = A_d J_s (e^{qV/kT} - 1)$$

Tunneling current:

Eq. (A.15) $$I_T = A_J \frac{q^3 E V}{4\pi^2 \hbar^2} \sqrt{\frac{m_e^*}{E_g}} \exp\left[-\frac{4\sqrt{2m_e^*} E_g^{3/2}}{3q\hbar E}\right]$$

Generation-recombination (G-R) current:

Eq. (A.16) $$I_{GR} = A_J \frac{2qn_i w}{\sqrt{\tau_{e0}\tau_{h0}}} \frac{kT \sinh\left(\frac{eV}{2kT}\right)}{q(V_{bi} - V)} f(b)$$

and *Surface leakage current*:

This current is the result of low resistance electrical paths along the sidewalls. It is typically proportional to the perimeter of the diodes as the leakage channels are confined with a few hundred nanometers from the sidewalls.

	p-n junction	NPJ
Area *p-n*	6.25 10^{-6} cm^2	10^{-10} cm^2
Perimeter *p-n*	100 μm	400 nm
Diffusion	1	?
G-R	1	1.6×10^{-5}
Tunnel	1	1.6×10^{-5}
Surface	1	<0.004

Table A.2. Comparison of the area, perimeter and the different components of the dark current for a regular p-n junction and a NPJ. The diffusion, G-R, tunnel and surface leakage were normalized to the regular p-n junction components for the comparison.

In the case of the NPJ, the effective area of the *p-n* junction is very small (10^{-10} cm^2) compared to regular etched *p-n* junctions (10^{-5} cm^2). Therefore, according to Eq. (A.15) and Eq. (A.16), the tunneling and G-R currents are five orders of magnitude smaller in NPJ diodes. As the perimeter is smaller too, the surface leakage should be reduced by almost three orders of magnitude. In addition, the surface of the junction is not damaged by plasma etching in this configuration. The variation of the diffusion current between the two structures is more complex to evaluate and would require more calculations. At the present time, the performance of InAs/GaSb devices is limited by the tunneling and surface leakage components of the dark current. Therefore, the NPJ structure should significantly reduce the dark current, thus improving the uniformity of the arrays.

However, this should not reduce the responsivity of the detectors because the optical area does not depend on the size of the electrical *p-n* junction. In the case of the NPJ, the entire π-region is used to absorb photons. This, added to the long diffusion length of the material, ensures that the quantum efficiency will not be decreased.

The NPJ should decrease the dark current without affecting the optical response. This would lead to higher signal to noise ratio and higher operability.

NPJ Device Fabrication

A slightly doped *p*-type VLWIR absorbing region is grown on the *p*-type GaSb substrate using molecular beam epitaxy. This layer will be used as the primary absorption layer. A silicon dioxide passivation/mask is then deposited using plasma enhanced chemical vapor deposition (PECVD). The sample is then coated with PMMA and small 100 nm openings are defined using e-beam lithography. The SiO_2 layer is then opened using inductively coupled plasma etching. The PMMA is then removed. The next step consists in implanting ions into the π-layer to create a local *n*-type material. Such local doping has never been performed on this material platform before. We propose to achieve it with one of the following techniques:

- Ion implantation
- Plasma immersion (in ICP)
- Diffusion of dopants in a MOCVD reactor
- Spin-on coating

Once the *p-n* junction has been performed, the rest of the processing is similar to the regular FPA processing. It is even possible to etch the material in between the unit cells to suppress any cross talk. The main difficulty comes from the local doping of the material.

Minority Electron Unipolar Photodetector (ME-UP)

Introduction

In semiconductor photon detectors, the absorption process results in a creation of an electron/hole pair which can be efficiently collected to produce an electrical signal. Based on the method to collect the photo-generated electron/hole pairs, the semiconductor detectors are then divided into two main categories: photoconductive and photovoltaic.

In a photoconductive detector, the photo-excited carriers change the conductance of the device (Fig. A. 20-top). Under a constant bias, the current difference between with and without illumination determines how efficiently the device operates. The advantage of the photoconductive mode is the simplicity in its operation and the flexibility in design architecture. Epitaxially grown photoconductive detector can be used in lateral configuration where the current flows horizontally along the sample and the incident radiation is illuminated vertically from the top (or the bottom through the substrate). In this configuration, the device area for optical absorption and electrical transport are decoupled, leading to efficient detection with a large optical area for increased photocurrent and a smaller electrical area for reduced noise current. However, this configuration is at the same time a drawback of photoconductor in arrayed multi-element detectors, due to the complexity of electrical collection. At small scale pixel, the ratio between the optical area over the electrical area is also reduced, leading to a lower signal to noise ratio. Due to the large dark current level, photoconductors are seldom used in vertically electrical collection mode (similar to that of photovoltaic detector) that is compatible for focal plane array fabrication. Only the family of low dimensional photoconductor (quantum well, quantum dots) can be used in vertical collection mode because they have relatively low dark currents due to quantum confinement of carriers, however, this quantum confinement exhibits a negative effect on the photocurrent, resulting in a low quantum efficiency.

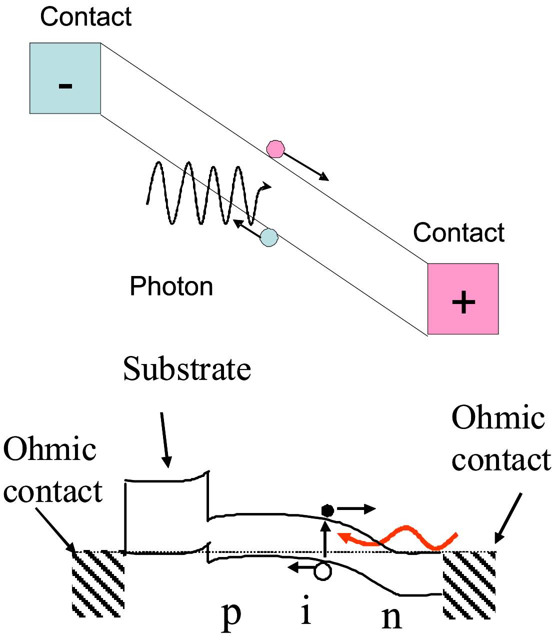


Fig. A. 20 Schematic diagram of the operation principle of two common types of photon detectors: top) photoconductive detector and bottom) photovoltaic detector.

In contrast, a photovoltaic detector is a *p-n* (or *p-i-n*) diode operating at the reverse bias where the electron/hole pairs are separated by the built-in field of the junction (Fig. A. 20-Bottom). Thanks to this driving force, photo-generated carriers can be efficiently collected at the electrodes with very small (or without) externally applied bias. One major advantage of photovoltaic detectors over photoconductors is the transport mechanisms of the “dark” carriers. While the transport in photoconductors involves majority carriers (which explains why photoconductors have high dark current), only transport of minority carriers are allowed in a negatively biased photodiode, due to a strong band bending of the *p*-*n* junction. Photovoltaic detectors have very low dark current level and do not require large ratio between optical area and electrical area. The current can flow along the growth direction, thus favoring the vertical electrical collection mode for arrayed detectors. Recent efforts in structural design have given birth to novel device architectures such as the heterostructures, double heterostructures, double layer planar heterostructures [Aries et al. 1993], graded band gap [Vurgaftman et al. 2006], *p-M-n* design [Nguyen et al. 2007a] etc., which have significantly enhanced the device performance. The common motivation of these architectures is to minimize the dark current due to the diffusion, generation-recombination and tunneling mechanisms. It turns out that the diffusion and generation-recombination components can be reduced by increasing the doping concentration at the active region. However, it would induce to the rise of the tunneling current due to an importance built-in electric field inside the depleted layer, especially at the long and very long wavelength infrared devices.

Both photodetector configurations have been widely utilized and exhibited their own advantages and disadvantages. In this section, we present a novel device architecture which can benefit from the advantages of both traditional configurations and can significantly enhance the device performance.

Device design

Using the type-II superlattice, the novel minority electron unipolar photodetector (ME-UP) architecture is shown in Fig. A. 21-Top. The device consists of two *p*-doped active regions and a thin barrier using the M-structure [Nguyen et al. 2007b] which has zero conduction band discontinuity with respect to the *p*-type active regions. Without the M-structure barrier, the device is a traditional photoconductor, in which the dark current is due to the transport of holes, the majority carriers in *p*-type conductors. With the presence of the M-structure, the layer acts as a potential barrier in the valence band, blocking all the holes' transport. The conductivity of the device is then based on the transport of the minority carriers: the electrons. To this end, the operation of the device is similar to that of a photovoltaic detector, and benefits from a very low dark current of diode-like devices. However, in this novel design, the absence of the charge transfer effect and of the depletion region will significantly alleviate the tunneling process. The *p*-type doping level in the active region can be increased to minimize the minority diffusion and generation-recombination without inducing a tunneling leakage. Moreover, in comparison with the standard photodiode configuration, the *p*-M-*p* device will experience much less generation-recombination and surface leakage current, thanks to the wide bandgap M-structure barrier. This will result to a significantly lower dark current and lower noise. The signal channel, however, is not affected by the M-structure since the photocurrent is collected in the conduction band, where no barrier is presence. Thus, in comparison with the traditional photovoltaic design, the ME-UP can exhibit much lower noise with lower dark current, while keeping the same optical efficiency.

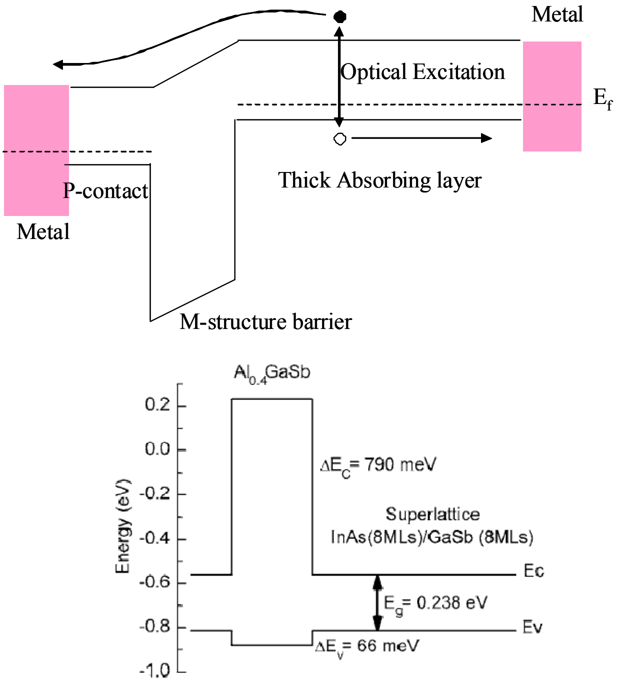

Metal
Optical Excitation
E_f
P-contact
Metal
Thick Absorbing layer
M-structure barrier
$Al_{0.4}GaSb$
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Energy (eV)
ΔE_C= 790 meV
Superlattice
InAs(8MLs)/GaSb (8MLs)
Ec
E_g= 0.238 eV
Ev
ΔE_V= 66 meV

Fig. A. 21. Schematic diagram of the photoconductive-photovoltaic hybrids: the novel ME-UP design (Top) shows many advantages over the pre-existed n-B-n design (Bottom).

Advantages of the ME-UP architecture

The ME-UP architecture belongs to the minority carrier photoconductor family, where the dark current is due to the transport of minority carrier. A similar design, called *n*-B-*n* structure [Maimon 2006]. (Fig. A. 21-Bottom), has been proposed and demonstrated the suppression of the generation-recombination and the surface leakage in InAs/AlAsSb MWIR detectors. The ME-UP design, in principle, shares the same advantages of the nBn design, which have already been demonstrated experimentally. Moreover, by using only superlattice-based materials, the ME-UP can be superior to the *n*-B-*n* design in both the device physics and practical realization.

First, InAs/GaSb multilayer materials, especially M-structure has demonstrated its great flexibility in engineering the electronic band structure. This will avoid the unexpected problem in nBn structure where the barrier is a bulk compound. Due to the monotonic relation between the compound composition and the energy gap, and the lattice constant, the barrier in the nBn structure could be a barrier at the valence band, which blocks the transport of photogenerated carriers, and reduce the device optical efficiency. Whereas in ME-UP design, thousands of possible design configurations will enable for the selection of a pair of active region/barrier with lattice matched condition and good band alignment.

All the layers of the ME-UP design have the same growth condition. This will avoid impractical growth difficulties that the nBn device encounters in the case where growth condition of the barrier is much different from the *n*-doped active region. The different growth conditions in nBn devices risks degrading the material quality as the top layer requires to be grown at a high temperature that could damage the layer underneath.

Thanks to the flexibility in controlling the band edge alignment, the ME-UP device can be designed to have a zero-band discontinuity in the conduction band. The active region can therefore be doped *p*-type. As opposed to the *n*-B-*n* design where the minority carriers are holes, the minority carriers in the ME-UP device are electrons. Since electrons have much longer diffusion length and higher mobility than hole, the collection of the photo-generated carrier is more efficient. Thus, the ME-UP design can potentially have higher quantum efficiency than the *n*-B-*n* design.

By using the *p*-type active region, the probability that a majority carrier tunnels through the barrier is much less than in the *n*-B-*n* design, since the effective mass of hole is much larger than that of electrons. Moreover, the M-structure has been shown to have large effective mass [Nguyen et al. 2008]. By using the M-structure as a barrier in ME-UP design, the barrier thickness can be small in comparison to the barrier thickness in the *n*-B-*n* design. Thus, in the case that there is minor band discontinuity in the conduction band of the ME-UP design, the electron photocurrent will be able to overcome the thin barrier with help of a small bias. There should not be significant bias dependence of the quantum efficiency like in the *n*-B-*n* structure.

References

Arias, J.M., Pasko, J.G., Zandian, M., Shin, S.H., Williams, G.M., Bubulac, L.O., DeWames, R.E., and Tennant, W.E., *Applied Physics Letters* Vol. 62, "Planar *p*-on-*n* HgCdTe heterostructure photovoltaic detectors," pp. 976, 1993.

Chu, M., Terterian, S., Walsh, D., Gurgenian, H.K., Mesropian, S., Rapp, R.J., and Holley, W.D., *Proceedings of the SPIE* Vol. 5783, "Recent progress on LWIR and VLWIR HgCdTe focal plane arrays," pp. 243, 2005.

Gilmore, A.S., Bangs, J., Gerrish, A., Stevens, A., and Starr, B., *Proceedings of the SPIE* Vol 5783, "Advancements in HgCdTe VLWIR materials," pp. 223, 2005.

Maimon, S., and Wicks, G.W., *Applied Physics Letters* Vol. 89, "nBn detector, an infrared detector with reduced dark current and higher operating temperature," pp. 151109, 2006.

Manissadjian, A., Tribolet, P., Destefanis, G., and De Borniol, E., *Proceedings of the SPIE* Vol. 5783, "Long wave HgCdTe staring arrays at Sofradir: from 9 mu m to 13+ mu m cut-offs for high performance applications," pp. 231, 2005.

Mohseni, H., Litvinov, V.I., and Razeghi, M., *Physical Review B: Condensed Matter and Materials Physics* Vol. 58, "Interface-induced suppression of the Auger recombination in type-II InAs/GaSb superlattices," pp. 15378, 1998.

Mohseni, H., Tahraoui, A., Wojkowski, J., Razeghi, M., Brown, G.J., Mitchel, W.C., and Park, Y.S., *Applied Physics Letters* Vol. 77, "Very long wavelength infrared type-II detectors operating at 80 K," pp. 1572, 2000.

Mohseni, H., Wei, Y., and Razeghi, M., *Proceedings of the SPIE* Vol. 4288, "High-performance type-II InAs/GaSb superlattice photodiodes," pp. 191, 2001.

Nguyen, B.M., Hoffman, D., Delaunay, P.Y., and Razeghi, M., *Applied Physics Letters* Vol. 91, "Dark current suppression in type II InAs/GaSb superlattice long wavelength infrared photodiodes with M-structure barrier,", pp. 163511, 2007a.

Nguyen, B.M., Razeghi, M., Nathan, V., and Brown, G.J., *Proceedings of the SPIE* Vol 6479, "Type-II M structure photodiodes: an alternative material design for mid-wave to long wavelength infrared regimes," pp. 64790S, 2007b.

Nguyen, B.M., Hoffman, D., Delaunay, P.Y., Huang, E.K.W., Razeghi, M., and Pellegrino, J., *Applied Physics Letters* Vol. 93, "Band edge tunability of M-structure for heterojunction design in Sb based type II superlattice photodiodes," pp. 163502, 2008.

Vurgaftman, I., Aifer, E.H., Canedy, C.L., Tischler, J.G., Meyer, J.R., Warner, J.H., Jackson, E.M., Hildebrandt, G., and Sullivan, G.J., *Applied Physics Letters* Vol. 89, "Graded band gap for dark-current suppression in long-wave infrared W-structured type-II superlattice photodiodes," pp. 121114, 2006

Youngdale, E.R., Meyer, J.R., Hoffman, C.A., Bartoli, F.J., Grein, C.H., Young, P.M., Ehrenreich, H., Miles, R.H., and Chow, D.H., *Applied Physics Letters* Vol. 64, "Auger lifetime enhancement in InAs-$Ga_{1-x}In_xSb$ superlattices," pp. 3160, 1994.

Zegrya, G.G., and Andreev, A.D., *Applied Physics Letters* Vol. 67, "Mechanism of suppression of Auger recombination processes in type-II heterostructures," pp. 2681, 1995.

A.8. Physical properties and safety information of metalorganics

Table A.3 and Table A.4 summarize some of the basic thermodynamic properties of metalorganic sources commonly used in MOCVD, including their chemical formula and abbreviation, boiling point, melting point, and the expression of their vapor pressure as a function of temperature.

Additional information on their other important physical properties is also provided for a number of important metalorganic sources, including diethylzinc (Table A.5), trimethylindium (Table A.6), triethylindium (Table A.7), trimethylgallium (Table A.8), and triethylgallium (Table A.9).

In the rest of this Appendix, general information about the safety of metalorganic compounds will be given. This will be helpful in developing safety and health procedures during their handling.

Chemical reactivity

Metalorganics catch fire if exposed to air, react violently with water and any compound containing active hydrogen, and may react vigorously with compounds containing oxygen or organic halide.

Stability

Metalorganics are stable when stored under a dry, inert atmosphere and away from heat.

Fire hazard

Metalorganics are spontaneously flammable in air and the products of combustion may be toxic. Metalorganics are pyrophoric by the paper char test used to gauge pyrophoricity for transportation classification purposes [Mudry 1975].

Firefighting technique

Protect against fire by strict adherence to safe operating procedures and proper equipment design. In case of fire, immediate action should be taken to confine it. All lines and equipment which could contribute to the fire should be shut off. As in any fire, prevent human exposure to fire, smoke or products of combustion. Evacuate non-essential personnel from the fire area.

The most effective fire extinguishing agent is dry chemical powder pressurized with nitrogen. Sand, vermiculite or carbon dioxide may be used. CAUTION: re-ignition may occur. DO NOT USE WATER, FOAM, CARBON TETRACHLORIDE OR CHLOROBROMOMETHANE extinguishing agents, as these materials react violently and/or liberate toxic fumes on contact with metalorganics.

When there is a potential for exposure to smoke, fumes or products of combustion, firefighters should wear full-face positive-pressure self-contained breathing apparatus or a positive-pressure supplied-air respirator with escape pack and impervious clothing including gloves, hoods, aluminized suits and rubber boots.

Human health

Metalorganics cause severe burns. Do not get in eyes, on skin or clothing.

Ingestion and inhalation. Because of the highly reactive nature of metalorganics with air and moisture, ingestion is unlikely.

Skin and eye contact. Metalorganics react immediately with moisture on the skin or in the eye to produce severe thermal and chemical burns.

First aid

If contact with metalorganics occurs, immediately initiate the recommended procedures below. Simultaneously contact a poison center, a physician, or the nearest hospital. Inform the person contacted of the type and extent of exposure, describe the victim’s symptoms, and follow the advice given.

Ingestion. Should metalorganics be swallowed, immediately give several glasses of water but do not induce vomiting. If vomiting does occur, give fluids again. Have a physician determine if condition of patient will permit induction of vomiting or evacuation of stomach. Do not give anything by mouth to an unconscious or convulsing person.

Skin contact. Under a safety shower, immediately flush all affected areas with large amounts of running water for at least 15 minutes. Remove contaminated clothing and shoes. Do not attempt to neutralize with chemical agents. Get medical attention immediately. Wash clothing before reuse.

Eye contact. Immediately flush the eyes with large quantities of running water for a minimum of 15 minutes. Hold the eyelids apart during the flushing to ensure rinsing of the entire surface of the eyes and lids with water. Do not attempt to neutralize with chemical agents. Obtain medical attention as soon as possible. Oils or ointments should not be used at this time. Continue the flushing for an additional 15 minutes if a physician is not immediately available.

Inhalation. Exposure to combustion products of this material may cause respiratory irritation or difficulty with breathing. If inhaled, remove to fresh air. If not breathing, clear the victim's airway and start mouth-to-mouth artificial respiration which may be supplemented by the use of a bag-mask respirator or manually triggered oxygen supply capable of delivering one liter per second or more. If the victim is breathing, oxygen may be delivered from a demand-type or continuous-flow inhaler, preferably with a physician's advice. Get medical attention immediately.

Industrial hygiene

Ingestion. As a matter of good industrial hygiene practice, food should be kept in a separate area away from the storage/use location. Smoking should be avoided in storage/use locations. Before eating, hands and face should be washed.

Skin contact. Skin contact must be prevented through the use of fire-retardant protective clothing during sampling or when disconnecting lines or opening connections. Recommended protection includes a full-face shield, impervious gloves, aluminized polyamide coat, hood and rubber boots. Safety showers –with quick-opening valves which that stay open– should be readily available in all areas where the material is handled or stored. Water should be supplied through insulated and heat-traced lines to prevent freeze-ups in cold weather.

Eye contact. Eye contact with liquid or aerosol must be prevented through the use of a full-face shield selected with regard for use-condition exposure potential. Eyewash fountains, or other means of washing the eyes with a gentle flow of tap water, should be readily available in all areas where this material is handled or stored. Water should be supplied through insulated and heat-traced lines to prevent freeze-ups in cold weather.

Inhalation. Metalorganics should be used in a tightly closed system. Use in an open (e.g. outdoor) or well ventilated area to minimize exposure to the products of combustion if a leak should occur. In the event of a leak, inhalation of fumes or reaction products must be prevented through the use of an approved organic vapor respirator with dust, mist, and fume filter. Where exposure potential necessitates a higher level of protection, use a positive-pressure, supplied-air respirator.

Spill handling

Make sure all personnel involved in spill handling follow proper firefighting techniques and good industrial hygiene practices. Any person entering an area with either a significant spill or an unknown concentration of fumes or combustion products should wear a positive-pressure, supplied-air respirator with escape pack. Block off the source of spill, extinguish fire with extinguishing agent. Re-ignition may occur. If the fire cannot be controlled with the extinguishing agent, keep a safe distance, protect adjacent property and allow product to burn until consumed.

Corrosivity to materials of construction

This material is not corrosive to steel, aluminum, brass, nickel or other common metals when blanketed with a dry inert gas. Some plastics and elastomers may be attacked.

Storage requirements

Containers should be stored in a cool, dry, well ventilated area. Store away from flammable materials and sources of heat and flame. Exercise due caution to prevent damage to or leakage from the container

Compound	Formula	Abbreviation	Melting point (°C)	Boiling point (°C)	$\mathrm{Log_{10}}\,P$(mmHg) (T in K)	Temperature Range (°C)
Group II sources						
Dimethylberyllium	$(CH_3)_2Be$	DMBe				
Diethylberyllium	$(C_2H_5)_2Be$	DEBe	12	194	$7.59 - 2200/T$	
Bis-cyclopentadienyl magnesium	$(C_5H_5)_2Mg$	Cp_2Mg	176		$25.14 - 2.18 \ln T - 4198/T$	
Group IIB sources						
Dimethylzinc	$(CH_3)_2Zn$	DMZn	-42	46	$7.802 - 1560/T$	
Diethylzinc	$(C_2H_5)_2Zn$	DEZn	-28	118	$8.280 - 2190/T$	
Dimethylcadmium	$(CH_3)_2Cd$	DMCd	-4.5	105.5	$7.764 - 1850/T$	
Group III sources						
Trimethylaluminum	$(CH_3)_3Al$	TMAl	15.4	126	$7.3147 - 1534.1/(T-53)$	17~100
Triethylaluminum	$(C_2H_5)_3Al$	TEAl	-58	194	$10.784 - 3625/T$	110~140
Trimethylgallium	$(CH_3)_3Ga$	TMGa	-15.8	55.7	$8.07 - 1703/T$	
Triethylgallium	$(C_2H_5)_3Ga$	TEGa	-82.3	143	$8.224 - 2222/T$	50~80
Ethyldimethylindium	$(CH_3)_2(C_2H_5)In$	EDMIn	5.5			10~38
Trimethylindium	$(CH_3)_3In$	TMIn	88.4	133.8	$10.520 - 3014/T$	
Triethylindium	$(C_2H_5)_3In$	TEIn	-32	184	1.2	44
					3	53
					12	83

Table A.3. Physical properties of some organometallics used in MOCVD. [Ludowise 1985, and http://electronicmaterials.rohmhaas.com]

Compound	Formula	Abbreviation	Melting point (°C)	Boiling point (°C)	$\text{Log}_{10} P$(mmHg) (T in K)	Temperature Range (°C)
Group IV sources						
Tetramethylgermanium	$(CH_3)_4Ge$	TMGe	-88	43.6	139	0
Tetramethyltin	$(CH_3)_4Sn$	TMSn	-53	78	$7.495 - 1620/T$	
Tetraethyltin	$(C_2H_5)_4Ge$	TESn	-112	181		
Group V sources						
Diethylarsine Hydride	$(C_2H_5)_2AsH$	DEAs			$7.339 - 1680/T$	
Tertiarybutylarsine	$(C_4H_9)AsH_2$	TBAs			$7.5 - 1562.3/T$	
Tertiarybutylphosphine	$(C_4H_9)PH_2$	TBP			$7.586 - 1539/T$	
Trimethylphosphorus	$(CH_3)_3P$	TMP	-85	37.8	$7.7329 - 1512/T$	
Triethylphosphorus	$(C_2H_5)_3P$	TEP	-88	127	$7.86 - 2000/T$	18~78.2
Trimethylarsenic	$(CH_3)_3As$	TMAs	-87.3	50~52	$7.7119 - 1563/T$	
Triethylarsenic	$(C_2H_5)_3As$	TEAs	-91	140	15.5	37
Trimethylantimony	$(CH_3)_3Sb$	TMSb	-86.7	80.6	$7.7280 - 1709/T$	
Triethylantimony	$(C_2H_5)_3Sb$	TESb	-98	116	17	75
Group VI sources						
Diethylselenide	$(C_2H_5)_2Se$	DESe	-	108		
Dimethyltellurium	$(CH_3)_2Te$	DMTe	10	82	$7.97 - 1865/T$	
Diethyltellurium	$(C_2H_5)_2Te$	DETe	-	137~138	$7.99 - 2093/T$	

Table A.4. Physical properties of some organometallics used in MOCVD. [Ludowise 1985, and http://electronicmaterials.rohmhaas.com]

Acronym	DEZn
Formula	$(C_2H_5)_2Zn$
Formula weight	123.49
Metallic purity	99.9999 wt% (min) zinc
Appearance	Clear, colorless liquid
Density	1.198 $g \cdot ml^{-1}$ at 30 °C
Melting point	−30 °C
Vapor pressure	3.6 mmHg at 0 °C 16 mmHg at 25 °C 760 mmHg at 117.6 °C
Behavior towards organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms relatively unstable complexes with simple ethers, thioethers, phosphines and arsines, but more stable complexes with tertiary amines and cyclic ethers.
Stability in air	Ignites on exposure (pyrophoric).
Stability in water	Reacts violently, evolving gaseous hydrocarbons, carbon dioxide and water.
Storage stability	Stable indefinitely at ambient temperatures when stored in an inert atmosphere.

Table A.5. Chemical properties of diethylzinc. [Razeghi 1989]

Acronym	TMIn
Formula	$(CH_3)_3In$
Formula weight	159.85
Metallic purity	99.999 wt% (min) indium
Appearance	White, crystalline solid
Density	1.586 g·ml^{-1} at 19 °C
Melting point	89 °C
Boiling point	135.8 °C at 760 mmHg 67 °C at 12 mmHg
Vapor pressure	15 mmHg at 41.7 °C
Stability in air	Pyrophoric, ignites spontaneously in air.
Solubility	Completely miscible with most common solvents.
Storage stability	Stable indefinitely when stored in an inert atmosphere.

Table A.6. Chemical properties of trimethylindium. [Razeghi 1989]

Acronym	TEIn
Formula	$(C_2H_5)_3In$
Formula weight	202.01
Metallic purity	99.9999 wt% (min) indium
Appearance	Clear, colorless liquid
Density	1.260 $g{\cdot}ml^{-1}$ at 20 °C
Melting point	−32 °C
Vapor pressure	1.18 mmHg at 40 °C 4.05 mmHg at 60 °C 12.0 mmHg at 80 °C
Behavior towards organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms complexes with ethers, thioethers, tertiary amines, phosphines, arsines and other Lewis bases.
Stability in air	Ignites on exposure (pyrophoric).
Stability in water	Partially hydrolyzed; loses one ethyl group with cold water.
Storage stability	Stable indefinitely at ambient temperatures when stored in an inert atmosphere.

Table A.7. Chemical properties of triethylindium. [Razeghi 1989]

Acronym	TMGa
Formula	$(CH_3)_3In$
Formula weight	114.82
Metallic purity	99.9999 wt% (min) gallium
Appearance	Clear, colorless liquid
Density	1.151 g·ml^{-1} at 15 °C
Melting point	−15.8 °C
Vapor pressure	64.5 mmHg at 0 °C 226.5 mmHg at 25 °C 760 mmHg at 55.8 °C
Behavior towards organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms complexes with ethers, thioethers, tertiary amines, tertiary phosphines, tertiary arsines, and other Lewis bases.
Stability in air	Ignites on exposure (pyrophoric).
Stability in water	Reacts violently, forming methane and Me_2GaOH or $[(Me_2Ga)_2O]_x$.
Storage stability	Stable indefinitely at ambient temperatures when stored in an inert atmosphere.

Table A.8. Chemical properties of trimethylgallium. [Razeghi 1989]

Acronym	TEGa
Formula	$(C_2H_5)_3Ga$
Formula weight	156.91
Metallic purity	99.9999 wt% (min) gallium
Appearance	Clear, colorless liquid
Density	1.0586 $g \cdot ml^{-1}$ at 20 °C
Melting point	−82.3 °C
Vapor pressure	16 mmHg at 43 °C 62 mmHg at 72 °C 760 mmHg at 143 °C
Behavior towards organic solvents	Completely miscible, without reaction, with aromatic and saturated aliphatic and alicyclic hydrocarbons. Forms complexes with ethers, thioethers, tertiary amines, tertiary phosphines, tertiary arsines and other Lewis bases.
Stability in air	Ignites on exposure (pyrophoric).
Stability in water	Reacts vigorously, forming ethane and Et_2GaOH or $[(Et_2Ga)_2O]_x$.
Storage stability	Stable indefinitely at room temperatures in an inert atmosphere.

Table A.9. Chemical properties of triethylgallium. [Razeghi 1989]

References

Ludowise, M., “Metalorganic chemical vapor deposition of III-V semiconductors,” *Journal of Applied Physics* 58, R31-R55, 1985.

Mudry, W.L., Burleson, D.C., Malpass, D.B., and Watson, S.C., *Journal of Fire Flammability* 6, p. 478, 1975.

Razeghi, M., *The MOCVD Challenge Volume 1: A Survey of GaInAsP-InP for Photonic and Electronic Applications*, Adam Hilger, Bristol, UK, 1989.

Sze, S.M., *Physics of Semiconductor Devices*, John Wiley & Sons, New York, 1981.

.

Index

1/f noise...330
2-color FPA..416
6.1 Angstrom family...........................368
abrupt junction....................................134
absorption...212
absorption region................................435
active avalanche quenching................441
active region..224
After-pulsing 446
alloying..117
amplification 175
angle-lap method..................................77
anti-reflectivity coating 293
APD..353, 426
Arrhenius..50
Auger recombination..................382, 525
Auger transition..................................372
avalanche multiplication.....................354
avalanche multiplication gain.............431
avalanche photodiode.................353, 426
background carrier concentration.........62
background limited infrared performance
....334
back-side illumination........................353
band bending 146
base transport factor 183
base-to-collector current amplification
factor 184
Bernard-Durafforg condition..............234
binary–binary SL................................372
biological agent detection...................460
bipolar laser..259
bipolar transistors 173
BJT...174
BLIP...334
bolometer..336
bound-to-bound 473
bound-to-continuum 473
Bragg grating......................................106
Bragg's law 250
breakdown voltage 429
Bridgman crystal growth 4
broad area laser...................................242
broken gap..369
built-in electric field...........................136
built-in potential.................................142
buried ridge 290
buried-heterostructure laser................245
carrier dynamics 276
cavity length.......................................307
channel 196
characterization 78
charge-coupled device........................363
chemical beam epitaxy 35
chirped superlattice 472
Chynoweth's law................................431
cladding layers 218
cleavage planes..................................232
coherent QDs.....................................406
collector...177
common base, transistor.....................188
common emitter, transistor................189
complementary error function 62
compositional fluxtuation...................384
confinement factor 224
constant-source diffusion 61
continuious wave...............................295
contrast...102
Cracker cell 31
critical electric field...........................429
critical thickness.................................248
crucible...120
crystal growth..1
current gain 183
current responsivity 326
current transfer ratio 184
cutoff frequency 223
cut-off wavelength 344
cyclotron orbit 474
Czochralski growth 3
dark count rate...................................445
dark counts 438
dark current 349, 399
dead time...446
Deal-Grove oxidation model................56
defect center, DX 192
depletion approximation.....135, 138, 155
depletion width..................137, 145, 164
detectivity..................................332, 333
detrapping time..................................446
DFB...250, 299
die chip...122

difference-frequency generation.468, 480
differential quantum efficiency 237
differential resistance 328, 350
diffusion 57
diffusion coefficient 59
diffusion current 149, 158
diode equation 155, 160
dip-pen lithography 97
direct patterning....................................93
distributed feedback 250, 299
dose 63, 70
dot-in-a-well (DWELL) 414
double channel....................................290
double heterostructure 254
double-heterojunction laser 240
double-heterostructure laser 240
drain 196
drift current................................151, 153
drive-in 61, 64
dry etching..110
dry oxidation method............................43
dual line package 127
duantum dot infrared photodetector....396
Early effect 185, 188
effective gain 440
effective mass 382
effective Richardson constant.....356, 522
Effusion cell 28
Einstein coefficients 213
electric displacement 217
electric field..217
electroepitaxy 10
electromagnetic radiation 323
electromagnetic spectrum...................457
electron transit time 347
electron-beam evaporation 120
electron-beam lithography.............95, 98
ellipsometry..53
emissivity-compensated pyrometry......23
emitter 177
emitter injection efficiency.................183
Empirical Tight Binding Model 378
epilayer-down bonding.......................304
epilayer-up bonding............................304
excess bias 438
excess noise...............................433, 440
exposure 91
extreme quantum limit........................476
extrinsic 343
Fabry Perot.........................229, 299, 308
far-field pattern...................................226
FET 196
fiber optic communications 252
Fick's first law......................................44
field effect transistor..................173, 196
filament evaporation...........................119
focal plane array. 127, 358, 363, 388, 533
forward biased *p-n* junction................148
forward scattering...............................100
four-point probe 75
FPA 363
free-carrier absorption 247
Fully Encapsulated Czochralski 4
gain...234
gain curve 215
gain-guided laser 245
Gas-source molecular beam epitaxy.....33
gated detection 442
Gauss's law 140
Gaussian 64
Geiger mode.......................426, 427, 436
generation-recombination noise 331
G-R noise 331
groove and stain method.......................77
HBT..190
heat sink 242
heat spreader 305
HEMT 203
Henry's law 46
heterojunction bipolar transistor.........190
heterojunction laser 239
high operating temperature FPA 412
high reflectivity coating 293
homojunction laser 239
hydride 14
ideal gas law 119
ideal *p-n* junction diode......................134
ideality factor 169
IMPact Avalanche Transit Time 464
impact ionization 353
implantation 57, 69
in situ characterization 20
in situ deflectometer.............................25
industrial hygiene 545
infrared 321
injection efficiency.............................283
interstitial diffusion 59
intersubband 255, 274

intersubband detector 396
intrinsic..343
ion implantation....................................69
ion milling 115
ion-beam lithography 95
ionization coefficients 354, 431, 432
ionization rate.....................................431
JFET 196
Johnson noise 329
junction depth................................63, 76
k·p theory..378
kinetic theory of gases........................119
laser 209
laser threshold 262
lattice mismatched..............................285
lattice-matched 240
LIDAR 355
lift-off technique...................................93
limited-source diffusion 61
Liquid Encapsulated Czochralski 3
Liquid phase electroepitaxy..................10
liquid phase epitaxy................................7
LO phonon scattering.................474, 475
load line..175
longitudinal optical modes 215
LPE 7
magnetic field.....................................474
magnetic field strength.......................217
magnetic induction 217
majority carriers 149
mask 84
mass transfer coefficient.......................45
Maxwell's equations...........................216
MBE..27
mean free path 119
melt epitaxy..9
mercury cell..32
MESFET 196
metal contact 163
metal interconnection 116
metallization.......................................116
metallurgic junction............................163
Metalorganic chemical vapor deposition
   14
Metalorganic MBE...............................35
metalorganic source......................15, 543
metal-oxide-semiconductor...........42, 54
metal-semiconductor junction ...134, 163, 168
metal-semiconductor-metal (MSM) ...357
metastable state 229
miniband 258, 275, 378
minority carrier extraction..................156
minority carrier injection....................156
minority carrier lifetime 525
minority carriers 149, 151
minority electron unipolar photodetector
   539
mirror loss 235, 281
misaligned 369
MISFET 196
MOCVD...14
MOCVD growth chamber....................18
modal threshold gain 248
molecular beam epitaxy...............27, 385
MOMBE 35
MOSFET...................................196, 200
MQW 247
M-structure superlatice.......................539
M-structure superlattice......................372
multiplication region 435, 436
multi-quantum well 247
nano-imprint lithography......................96
Nano-Planar Junction 535
noise 326
noise equivalent circuit.......................328
noise spectral density 327
noise-equivalent-power 332
nominal dose 102
normal incidence 402
ohmic contact 167
optical confinement factor..................279
Optical Parametric Amplifier 530
optical parametric oscillator 467
oscillator strength 402
output power scaling 306
overbias 438
oxidation rate..50
oxide film color 54
ozone 451
packaging 122, 290
passivation..387
passive avalanche quenching.............441
passive quenching 437
PCDFB 308, 312
periodic table 503
permeability 217
permittivity...217

phonon bottleneck 400, 407
photoconductive decay.......................529
photoconductive detector 345
photoconductive gain 347, 398
photoconductivity...............................345
photoconductor...................................345
Photoconductor modeling...................527
photocurrent 326, 346, 349
photodetectors 322
photodiode..348
Photodiode modeling..........................527
photoelectromagnetic (PEM) effect....360
photolithography 84, 89
photomask 89
photomultiplier tubes..........................425
photon detectors 343
photon noise 332
photonic crystal 308
photonic crystal distributed feedback 308, 312
photoresist 53, 85, 86
photoresponse............................344, 347
photovoltage.......................................326
photovoltaic detector..........................348
p-i-n photodiode 351
pinch-off.............................196, 197, 198
PIV 292
Planck's law 325
plasma etching...................................110
PMMA 100
p-n junction 65, 133
polar optical phonon..........................284
population inversion...214, 230, 233, 277
power conversion efficiency..............211
power-current-voltage characteristics.292
precursors...16
predeposition..................................61, 62
projected range 71
pseudo potential model.......................378
pulse detection efficiency..................444
pulse height discriminator 437
pulsed mode 295
pyroelectric...336
pyrometry 23
pyrophoric 543
QCL..272
QD cap layer 410
QDIP absorption coefficient..............409
QDIPs..396
Q-switching.......................................231
quantum cascade laser........255, 272, 472
quantum dot infrared photodetector ..362, 400
quantum dots.....................................262
quantum efficiency.....236, 281, 337, 346
quantum key distribution....................449
quantum size effects...........................404
quantum well.....................................247
quantum well infrared photodetector..396
quantum well intersubband photodetector 344, 361
quasi-equilibrium 234
quasi-Fermi energy.............................234
QW...247
QWIP 361, 396
radiative recombination lifetime.........212
Raman 264
rate equation model 276
rate equations 277
Reach-through avalanche photodiodes 435
reactive ion etching 114
read-out integrated circuit 363, 389
recessed window 429
recombination....................................193
rectifying contact................................166
Reflection anisotropy spectroscopy......21
reflection high energy electron diffraction 28
reflectomety 22
refractive index..................................215
resist 84
resonant cavity 215
resonant tunneling filter 413
responsivity 326, 347
reverse biased *p-n* junction.........151, 153
Richardson constant 169, 522
ROIC 363
ruby laser...228
saturation current................160, 330, 523
SCH...246, 252
Schottky barrier photodiodes.............355
Schottky contact.................166, 167, 168
Schottky potential barrier height 169, 356
scribing...122
secondary ion mass spectroscopy.........78
self-assembly.....................................405
semiconductor heterostructures............57

semiconductor laser....................211, 232
sensitivity 102
separate absorption chanrge
  multiplication 436
separate confinement heterostructure 246, 252
series resistance 237
sheet resistivity.....................................74
Shockley-Read-Hall recombination....525
short-circuit current 349
shot noise...330
signal-to-noise ratio....................327, 332
SIMS 78
single electron transistor.....................203
single-photon avalanche diodes..........437
single-photon counting module 448
single-photon detection 426
single-photon detection efficiency......444
slope efficiency 237, 281, 296
SLS...254
solid solubility 62
source 196
space charge region 137, 154
specific detectivity..............................333
spontaneous emission 212
sputter etching 114
sputtering deposition 121
step junction 134
stimulated emission 212, 279
straggle 72
strain-balanced 286
strained layer superlattices 254
Stranski-Krastanov growth.........362, 405
submount 304
substitutional diffusion 59
SUMO cell 30
superlattice 258
surface leakage current.......................387
surface-emitting lasers........................260
switching 175
TE polarization 219
TEC 292
temperature noise 331
terahertz...457
terminals..175
thermal conductance...........................337
thermal detector..................................336
thermal oxidation...........................42, 43
thermal packaging 304
thermal response time.........................338
thermionic emission 356, 401, 521
thermocouple......................................336
thermoelectric cooler.................249, 292
thermopile 336
threshold current density 236, 297
threshold, laser 234
THz laser 469
THz photomixer 466
THz sources, broadband 464
THz spectroscopy.......................458, 462
THz time domain spectroscopy 464
time correlated single-photon counting
 442
time-resolved photoluminescence 529
TM polarization..................................219
TO-style package 126
transparency gain 247
transparency point 234
transverse modes 216
trap-assisted tunneling................387, 429
T-ray imaging.....................................460
tunneling leakage current 383
turn-on voltage 237
type I 257
type II 257, 357
type II broken gap 370
type II misaligned...............................257
type II staggered 257
type II superlattice...... 257, 367, 381, 534
ultraviolet 322
unipolar device 396
unipolar laser......................................259
up-conversion.....................................448
UV sensing 451
vacuum deposition 118
van der Pauw 75
vapor phase epitaxy 11
Vapor pressure controlled Czochralski...4
vertical Bridgman...................................5
vertical cavity surface-emitting lasers 260
voltage responsivity............................326
VPE 11
wall plug efficiency............292, 296, 300
wave equation.....................................218
waveguide 216, 218, 228, 288, 301
wavevector 220
wet chemical etching..........................107
wet oxidation method 43

wetting layer.......................................405
white noise 328
Wien’s law 325
wire bonding.......................................123
work function 165
WPE 292, 296, 300
W-structure superlattice 371
x-ray lithography..................................95
Zener breakdown................................190
Zener tunneling 429
zero-read noise 440